MODELING PERFORMANCE MEASUREMENT

Applications and Implementation Issues in DEA

Wade D. Cook
Joe Zhu

DEAFrontier Software Included!

Springer

MODELING PERFORMANCE MEASUREMENT
Applications and Implementation Issues in DEA

MODELING PERFORMANCE MEASUREMENT
Applications and Implementation Issues in DEA

by

Wade D. Cook
York University, Canada

Joe Zhu
Worcester Polytechnic Institute, U.S.A.

Springer

Wade D. Cook
York University
Canada

Joe Zhu
Worcester Polytechnic Institute
USA

Library of Congress Cataloging-in-Publication Data
A C.I.P. Catalogue record for this book is available
from the Library of Congress.

MODELING PERFORMANCE MEASUREMENT: Applications and Implementation Issues in DEA

ISBN 0-387-24137-X e-ISBN 0-387-24138-8

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 11052760

springeronline.com

To the Memory
of
Marsha Jeanne Cook

CONTENTS

Chapter 10 Resource Constrained DEA

Chapter 11 Multicomponent Efficiency: Measurement and Core Business Identification in Multiplant Firms

Chapter 12 Implementation of Robotics: Identifying Efficient Implementors

Chapter 16 Context-Dependent DEA: Models and Extension

Chapter 17 Evaluating Power Plant Efficiency: Hierarchical Models

PREFACE

Data Envelopment Analysis (DEA) is a data-oriented approach for performance evaluation and improvement. In recent years, we have observed a notable increase in interest in DEA techniques and their applications. Basic DEA models and techniques have been well documented in the DEA literature. Although these basic DEA models are useful in determining the best-practice frontier, identification of best-practices is seldom the ultimate goal with respect to performance evaluation. It is generally important to further analyze the business operations after the identification of best-practice, so that in-depth managerial information can be derived. It is also important to correctly design and model the performance issues. Because of the complexity of the business or engineering operations which are often characterized by multiple functions, multiple stages and multiple levels, new (and advanced) DEA methods are needed to reconcile the multidimensional aspects of performance evaluation issues.

The book presents unified results from the authors' recent DEA research. New methodologies and techniques are developed in application-driven scenarios, to go beyond identification of the best-practice frontier, and seek solutions to aid managerial decisions. These new DEA developments are deeply grounded in real-world applications. DEA researchers and practitioners alike will find this book helpful. Theory is provided for DEA researchers for further development and possible extensions. However, each theory is also presented in a practical way for DEA practitioners via numerical examples, simple real management cases and verbal descriptions.

The book covers pure DEA applications in such areas as highway maintenance, technology implementations, and others. DEA methodology enhancements are wrapped into applications. New DEA theoretical developments are included, for example, on how to use DEA as a benchmarking tool, and how to use DEA in multi-criteria decision making. The book provides a balanced coverage of DEA for both academic researchers and industry practitioners. It addresses advanced/new DEA methodology and techniques that are developed for modeling unique and new performance evaluation issues. Some of the DEA models can be computed using the accompanying DEAFrontier software which is an Excel Add-In.

Wade D. Cook
Schulich School of Business
York University
4700 Keele Street
Toronto, Ontario
Canada M3J 1P3
Email: wcook@schulich.yorku.ca

Joe Zhu
Department of Management
Worcester Polytechnic Institute
Worcester, MA 01609 USA
Email: jzhu@wpi.edu

Chapter 1

DATA ENVELOPMENT ANALYSIS

1.1. INTRODUCTION

Data Envelopment Analysis (DEA) is a relatively new "data oriented" approach for evaluating the performance of a set of peer entities called Decision Making Units (DMUs) which convert multiple inputs to multiple outputs. The definition of a DMU is generic and flexible. Recent years have seen a great variety of applications of DEA for use in evaluating the performances of many different kinds of entities engaged in many different activities in many different contexts in many different countries. These DEA applications have used DMUs of various forms to evaluate the performance of entities, such as hospitals, US Air Force wings, universities, cities, courts, business firms, and others, including the performance of countries, regions, etc. Because it requires very few assumptions, DEA has also opened up possibilities for use in cases which have been resistant to other approaches because of the complex (often unknown) nature of the relations between the multiple inputs and multiple outputs involved in DMUs (Cooper, Seiford and Zhu, 2004).

Since DEA in its present form was first introduced in 1978, researchers in a number of fields have quickly recognized that it is an excellent and easily used methodology for modeling operational processes for performance evaluations (Cooper, Seiford and Tone, 2000). This has been accompanied by other developments. For instance, Zhu (2002) provides a number of DEA spreadsheet models that can be used in performance evaluation and benchmarking. DEA's empirical orientation and the absence of a need for the numerous *a priori* assumptions that accompany other approaches (such as standard forms of statistical regression analysis) have resulted in its use in a number of studies involving efficient frontier estimation in the governmental and nonprofit sector, in the regulated sector, and in the private sector.

In their originating study, Charnes, Cooper, and Rhodes (1978) described DEA as a 'mathematical programming model applied to observational data [that] provides a new way of obtaining empirical estimates of relations - such as the production functions and/or efficient production possibility surfaces – that are cornerstones of modern economics'.

Formally, DEA is a methodology directed to frontiers rather than central tendencies. Instead of trying to fit a regression plane through the *center* of the data as in statistical regression, for example, one 'floats' a piecewise linear surface to rest on top of the observations. Because of this perspective, DEA proves particularly adept at uncovering relationships that remain hidden from other methodologies. For instance, consider what one wants to mean by "efficiency", or more generally, what one wants to mean by saying that one DMU is more efficient than another DMU. This is accomplished in a straightforward manner by DEA without requiring expectations and variations with various types of models such as in linear and nonlinear regression models.

1.2. ENVELOPMENT AND MULTIPLIER DEA MODELS

Consider a set of n observations on the DMUs. Each observation, DMU_j ($j = 1, \ldots, n$), uses m inputs x_{ij} ($i = 1, 2, \ldots, m$) to produce s outputs y_{rj} ($r = 1, 2, \ldots, s$). The CCR ratio model can be expressed as

$$\max h_o(u,v) = \textstyle\sum_r u_r y_{ro} / \sum_i v_i x_{io} \tag{1.1}$$

where the variables are the u_r's and the v_i's and the y_{ro}'s and x_{io}'s are the observed output and input values, respectively, of $\boldsymbol{DMU_o}$, the DMU to be evaluated. Of course, without further additional constraints (developed below) (1.1) is unbounded.

A set of normalizing constraints (one for each DMU) reflects the condition that the virtual output to virtual input ratio of every DMU, including $DMU_j = DMU_o$, must be less than or equal to unity. The mathematical programming problem may thus be stated as

$$\begin{aligned} &\max h_o(u,v) = \textstyle\sum_r u_r y_{ro} / \sum_i v_i x_{io} \\ &\text{subject to:} \\ &\textstyle\sum_r u_r y_{rj} / \sum_i v_i x_{ij} \le 1 \text{ for } j = 1, \ldots, n, \\ &u_r, v_i \ge 0. \end{aligned} \tag{1.2}$$

The above ratio form yields an infinite number of solutions; if (u^*, v^*) is optimal, then (αu^*, αv^*) is also optimal for $\alpha > 0$. However, the transformation developed by Charnes and Cooper (1962) for linear fractional programming selects a representative solution [i.e., the solution (u, v) for which $\sum_{i=1}^{m} v_i x_{io} = 1$] and yields the equivalent linear programming problem [the change of variables from (u, v) to (μ, v) is a result of the Charnes-Cooper transformation],

$$\max z = \sum_{r=1}^{s} \mu_r y_{ro}$$

$$\text{subject to}$$

$$\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0 \qquad (1.3)$$

$$\sum_{i=1}^{m} v_i x_{io} = 1$$

$$\mu_r, v_i \ge 0$$

The dual program of (1.3) is

$$\theta^* = \min \theta$$

$$\text{subject to}$$

$$\sum_{j=1}^{n} x_{ij} \lambda_j \le \theta x_{io} \qquad i = 1,2,\dots,m; \qquad (1.4)$$

$$\sum_{j=1}^{n} y_{rj} \lambda_j \ge y_{ro} \qquad r = 1,2,\dots,s;$$

$$\lambda_j \ge 0 \qquad j = 1,2,\dots,n.$$

Since $\theta = 1$ is a feasible solution to (1.4), the optimal value to (1.4), $\theta^* \le 1$. If $\theta^* = 1$, then the current input levels cannot be reduced (proportionally), indicating that DMU_o is on the frontier. Otherwise, if $\theta^* < 1$, then DMU_o is dominated by the frontier. θ^* represents the (input-oriented) efficiency score of DMU_o.

Table 1-1. Supply Chain Operations Within a Week

DMU	Cost ($100)	Response time (days)	Profit ($1,000)
1	1	5	2
2	2	2	2
3	4	1	2
4	6	1	2
5	4	4	2

We now consider a simple numerical example shown in Table 1.1 where we have five DMUs (supply chain operations). Within a week, each DMU generates the same profit of $2,000 with a different combination of supply chain cost and response time.

Figure 1-1 presents the five DMUs and the piecewise linear frontier. DMUs 1, 2, 3, and 4 are on the frontier. If we calculate model (1.4) for DMU5,

$$
\begin{aligned}
&\text{Min } \theta \\
&\text{Subject to:} \\
&1\lambda_1 + 2\lambda_2 + 4\lambda_3 + 6\lambda_4 + 4\lambda_5 \leq 4\theta \\
&5\lambda_1 + 2\lambda_2 + 1\lambda_3 + 1\lambda_4 + 4\lambda_5 \leq 4\theta \\
&2\lambda_1 + 2\lambda_2 + 2\lambda_3 + 2\lambda_4 + 2\lambda_5 \geq 2 \\
&\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 \geq 0
\end{aligned}
$$

we obtain a set of unique optimal solutions of $\theta^* = 0.5$, $\lambda_2^* = 1$, and $\lambda_j^* = 0$ ($j \neq 2$), indicating that DMU2 is the benchmark for DMU5, and DMU5 should reduce its cost and response time to the amounts used by DMU2.

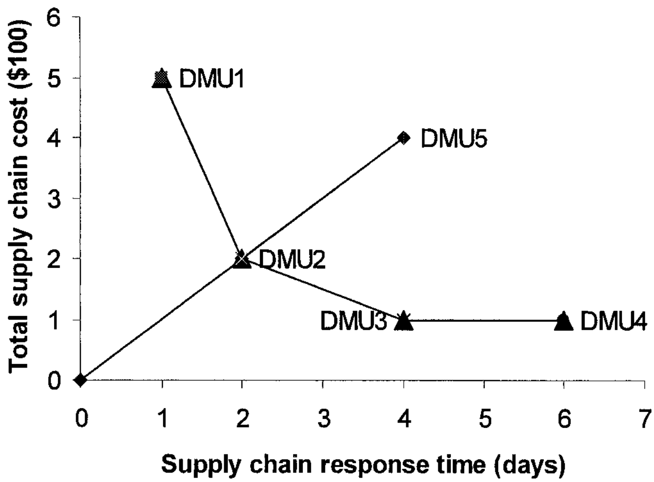

Figure 1-1. Five Supply Chain Operations

Now, if we calculate model (1.4) for DMU4, we obtain $\theta^* = 1$, $\lambda_4^* = 1$, and $\lambda_j^* = 0$ ($j \neq 4$), indicating that DMU4 is on the frontier. However, Figure 1-1 indicates that DMU4 can still reduce its response time by 2 days to reach DMU3. This individual input reduction is called input slack.

In fact, both input and output slack values may exist in model (1.4). Usually, after calculating (1.4), we have

$$\begin{cases} s_i^- = \theta^* x_{io} - \sum_{j=1}^{n} \lambda_j x_{ij} & i = 1,2,...,m \\ s_r^+ = \sum_{j=1}^{n} \lambda_j y_{rj} - y_{ro} & r = 1,2,...,s \end{cases} \tag{1.5}$$

where s_i^- and s_r^+ represent input and output slacks, respectively. An alternate optimal solution of θ^* = 1 and λ_3^* = 1 exists when we calculate model (1.4) for DMU4. This leads to s_1^- = 2 for DMU4. However, if we obtain θ^* = 1 and λ_4^* = 1 from model (1.4), we have all zero slack values. i.e., because of possible multiple optimal solutions, (1.4) may not yield all the non-zero slacks.

Therefore, we use the following linear programming model to determine the possible non-zero slacks after (1.2) is solved.

$$\max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

subject to

$$\begin{aligned} &\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta^* x_{io} && i = 1,2,...,m; \\ &\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = y_{ro} && r = 1,2,...,s; \\ &\lambda_j \geq 0 && j = 1,2,...,n. \end{aligned} \tag{1.6}$$

For example, applying (1.6) to DMU4 yields

$$\begin{aligned} &\text{Max } s_1^- + s_2^- + s_1^+ \\ &\text{Subject to} \\ &1\lambda_1 + 2\lambda_2 + 4\lambda_3 + 6\lambda_4 + 4\lambda_5 + s_1^- = 6\,\theta^* = 6 \\ &5\lambda_1 + 2\lambda_2 + 1\lambda_3 + 1\lambda_4 + 4\lambda_5 + s_2^- = 1\,\theta^* = 1 \\ &2\lambda_1 + 2\lambda_2 + 2\lambda_3 + 2\lambda_4 + 2\lambda_5 - s_1^+ = 2 \\ &\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, s_1^-, s_2^-, s_1^+ \geq 0 \end{aligned}$$

with optimal slacks of $s_1^{-*} = 2$, $s_2^{-*} = s_1^{+*} = 0$.

Definition 1.1 (DEA Efficiency): The performance of DMU_o is fully (100%) efficient if and only if both (i) θ^* = 1 and (ii) all slacks $s_i^{-*} = s_r^{+*} = 0$.

Definition 1.2 (Weakly DEA Efficient): The performance of DMU_o is weakly efficient if and only if both (i) θ^* = 1 and (ii) $s_i^{-*} \neq 0$ and/or $s_r^{+*} \neq 0$ for some *i* and *r*.

In Figure 1.2, DMUs 1, 2, and 3 are efficient, and DMU 4 is weakly efficient. (The slacks obtained by (1.6) are called DEA slacks.)

In fact, models (1.4) and (1.6) represent a two-stage DEA process involved in the following DEA model.

$$\min \theta - \varepsilon(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+)$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta x_{io} \qquad i = 1,2,...,m; \tag{1.7}$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{ro} \qquad r = 1,2,...,s;$$

$$\lambda_j \geq 0 \qquad j = 1,2,...,n.$$

The presence of the non-Archimedean ε in the objective function of (1.7) effectively allows the minimization over θ to preempt the optimization involving the slacks, s_i^- and s_r^+. Thus, (1.7) is calculated in a two-stage process with maximal reduction of inputs being achieved first, via the optimal θ^* in (1.4); then, in the second stage, movement onto the efficient frontier is achieved via optimizing the slack variables in (1.6).

In fact, the presence of weakly efficient DMUs is the cause of multiple optimal solutions. Thus, if weakly efficient DMUs are not present, the second stage calculation (1.6) is not necessary, and we can obtain the slacks using (1.5). However, priori to calculation, we usually do not know whether weakly efficient DMUs are present.

Model (1.7) is usually called "envelopment" DEA model. The dual program to (1.7) is called "multiplier" DEA model.

$$\max z = \sum_{r=1}^{s} \mu_r y_{ro}$$

subject to

$$\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0 \tag{1.8}$$

$$\sum_{i=1}^{m} v_i x_{io} = 1$$

$$\mu_r, v_i \geq \varepsilon > 0$$

If we consider the following DEA model,

$$\begin{aligned}&\text{Min } \textstyle\sum_i v_i x_{io} / \sum_r u_r y_{ro}\\&\text{Subject to}\\&\textstyle\sum_i v_i x_{ij} / \sum_r u_r y_{rj} \geq 1 \text{ for } j = 1, \ldots, n,\\&u_r,\, v_i \geq \varepsilon > 0.\end{aligned} \tag{1.9}$$

where $\varepsilon > 0$ is the previously defined non-Archimedean element, then we have the following output-oriented multiplier and envelopment DEA models

$$\begin{aligned}&\min q = \sum_{i=1}^{m} v_i x_{io}\\&\text{subject to}\\&\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} \geq 0\\&\sum_{r=1}^{s} \mu_r y_{ro} = 1\\&\mu_r, v_i \geq \varepsilon, \quad \forall r, i\end{aligned} \tag{1.10}$$

$$\begin{aligned}&\max \phi + \varepsilon\Big(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\Big)\\&\text{subject to:}\\&\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = x_{io} \quad i = 1,2,\ldots,m;\\&\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = \phi y_{ro} \quad r = 1,2,\ldots,s;\\&\lambda_j \geq 0 \qquad\qquad j = 1,2,\ldots,n.\end{aligned} \tag{1.11}$$

As before, model (1.11) is calculated in a two-stage process. First, we calculate ϕ^* by ignoring the slacks. Then we optimize the slacks by fixing ϕ^* in the following linear programming problem,

$$\max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

subject to

$$\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = x_{io} \qquad i = 1,2,...,m; \qquad (1.12)$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = \phi^* y_{ro} \qquad r = 1,2,...,s;$$

$$\lambda_j \geq 0 \qquad j = 1,2,...,n.$$

We then modify the previous input-oriented definition of DEA efficiency to the following output-oriented version.

Definition 1.3: DMU_o is efficient if and only if $\phi^* = 1$ and $s_i^{-*} = s_r^{+*} = 0$ for all i and r. DMU_o is weakly efficient if $\phi^* = 1$ and $s_i^{-*} \neq 0$ and (or) $s_r^{+*} \neq 0$ for some i and r.

The frontier determined by the above DEA models exhibits constant returns to scale (CRS). Thus, the above DEA models are called CRS DEA models with different orientations. Figure 1-2 shows a CRS frontier – ray OB. Based upon this CRS frontier, only B is efficient.

The constraint on $\sum_{j=1}^{n} \lambda_j$ in the envelopment models actually determines the returns to scale (RTS) type of an efficient frontier. If we add $\sum_{j=1}^{n} \lambda_j = 1$, we obtain VRS (variable RTS) models. The frontier is ABCD as shown in Figure 1-2.

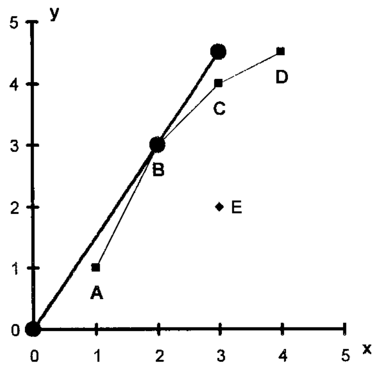

Figure 1-2. CRS Frontier

If we replace $\sum_{j=1}^{n}\lambda_j = 1$ with $\sum_{j=1}^{n}\lambda_j \leq 1$, then we obtain non-increasing RTS (NIRS) envelopment models. In Figure 1-3, the NIRS frontier consists of DMUs B, C, D and the origin.

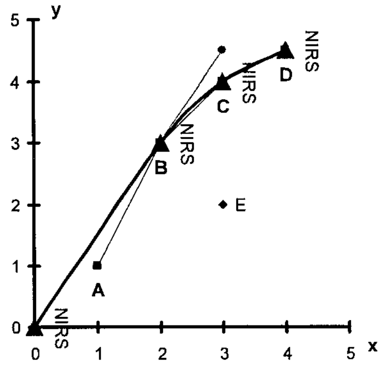

Figure 1-3. NIRS Frontier

If we replace $\sum_{j=1}^{n}\lambda_j = 1$ with $\sum_{j=1}^{n}\lambda_j \geq 1$, then we obtain non-decreasing RTS (NDRS) envelopment models. In Figure 1-4, the NDRS frontier consists of DMUs, A, B, and the section starting with B on ray OB.

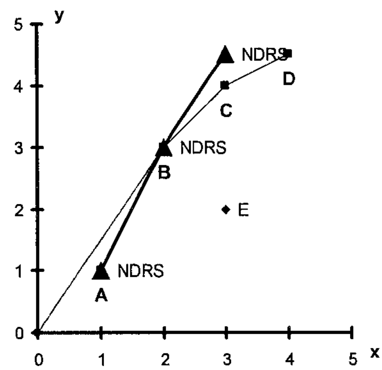

Figure 1-4. NDRS Frontier

Table 1-2 summarizes the envelopment and the multiplier models with respect to the orientations and frontier types. The last row presents the efficient target (DEA projection) of a specific DMU under evaluation.

Table 1-2. DEA Models

Frontier Type	Input-Oriented	Output-Oriented
CRS	$\min \theta - \varepsilon(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+)$ subject to $\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta x_{io} \quad i = 1,2,\ldots,m;$ $\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{ro} \quad r = 1,2,\ldots,s;$ $\lambda_j \geq 0 \quad j = 1,2,\ldots,n.$	$\max \phi + \varepsilon(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+)$ subject to $\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io} \quad i = 1,2,\ldots,m;$ $\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = \phi y_{ro} \quad r = 1,2,\ldots,s;$ $\lambda_j \geq 0 \quad j = 1,2,\ldots,n.$
VRS	Add $\sum_{j=1}^{n} \lambda_j = 1$	
NIRS	Add $\sum_{j=1}^{n} \lambda_j \leq 1$	
NDRS	Add $\sum_{j=1}^{n} \lambda_j \geq 1$	
Efficient Target	$\begin{cases} \hat{x}_{io} = \theta^* x_{io} - s_i^{-*} & i = 1,2,\ldots,m \\ \hat{y}_{ro} = y_{ro} + s_r^{+*} & r = 1,2,\ldots,s \end{cases}$	$\begin{cases} \hat{x}_{io} = x_{io} - s_i^{-*} & i = 1,2,\ldots,m \\ \hat{y}_{ro} = \phi^* y_{ro} + s_r^{+*} & r = 1,2,\ldots,s \end{cases}$

	$\max \sum_{r=1}^{s} \mu_r y_{ro} + \mu$ subject to $\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} + \mu \leq 0$ $\sum_{i=1}^{m} v_i x_{io} = 1$ $\mu_r, v_i \geq 0(\varepsilon)$	$\min \sum_{i=1}^{m} v_i x_{io} + v$ subject to $\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} + v \geq 0$ $\sum_{r=1}^{s} \mu_r y_{ro} = 1$ $\mu_r, v_i \geq 0(\varepsilon)$
CRS	where $\mu = 0$	where $v = 0$
VRS	where μ free	where v free
NIRS	where $\mu \leq 0$	where $v \geq 0$
NDRS	where $\mu \geq 0$	where $v \leq 0$

1.3. ASSURANCE REGION DEA MODELS

Note that the only restriction on the multiplier DEA models is the positivity of the multipliers imposed by ε. In the DEA literature, a number of approaches have been proposed to introduce additional restrictions on the values that the multipliers can assume.

Some of the techniques for enforcing these additional restrictions include imposing upper and lower bounds on individual multipliers (Dyson and Thanassoulis, 1988; Roll, Cook, and Golany, 1991); imposing bounds on ratios of multipliers (Thompson et al., 1986); appending multiplier inequalities (Wong and Beasley, 1990); and requiring multipliers to belong to given closed cones (Charnes et al., 1989).

We here present the assurance region (AR) approach of Thompson et al. (1986). To illustrate the AR approach, suppose we wish to incorporate additional inequality constraints of the following form into the multiplier DEA models as given in Table 1-2:

$$\begin{aligned} \alpha_i \le \frac{v_i}{v_{i_o}} \le \beta_i, \qquad & i = 1,\ldots,m \\ \delta_r \le \frac{\mu_r}{\mu_{r_o}} \le \gamma_r, \qquad & r = 1,\ldots,s \end{aligned} \tag{1.13}$$

Here, v_{i_o} and μ_{r_o} represent multipliers which serve as "numeraires" in establishing the upper and lower bounds represented here by α_i, β_i, and by δ_r, γ_r for the multipliers associated with inputs i =1, ..., m and outputs r = 1, ..., s where $\alpha_{i_o} = \beta_{i_o} = \delta_{r_o} = \gamma_{r_o} = 1$. The above constraints are called Assurance Region (AR) constraints as developed by Thompson et al. (1986) and defined more precisely in Thompson et al. (1990).

Uses of such bounds are not restricted to prices. For example, Zhu (1996a) uses an assurance region approach to establish bounds on the weights obtained from uses of Analytic Hierarchy Processes in Chinese textile manufacturing in order to reflect how the local government in measuring the textile manufacturing performance.

The generality of these AR constraints provides flexibility in use. Prices, utils and other measures may be accommodated and so can mixtures of such concepts. Moreover, one can first examine provisional solutions and then tighten or loosen the bounds until one or more solutions is attained that appears to be reasonably satisfactory to decision makers who cannot state the values for their preferences in an a priori manner.

1.4. SLACK BASED DEA MODELS

The input-oriented DEA models consider the possible (proportional) input reductions while maintaining the current levels of outputs. The output-oriented DEA models consider the possible (proportional) output augmentations while keeping the current levels of inputs. Charnes, Cooper, Golany, Seiford and Stutz (1985) develop an additive DEA model which considers possible input decreases as well as output increases simultaneously. The additive model is based upon input and output slacks. For example,

$$\max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$
$$\text{subject to:}$$
$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io} \qquad i = 1,2,...,m; \qquad (1.14)$$
$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{ro} \qquad r = 1,2,...,s;$$
$$\lambda_j, s_i^-, s_r^+ \geq 0$$

Note that model (1.8) assumes equal marginal worth for the nonzero input and output slacks. Therefore, caution should be excised in selecting the units for different input and output measures. Some *a priori* information may be required to prevent an inappropriate summation of non-commensurable measures. Previous management experience and expert opinion, which prove important in productivity analysis, may be used (see Seiford and Zhu (1998)).

Model (1.8) therefore is modified to a weighted CRS slack-based model as follows (Ali, Lerme and Seiford, 1995; Thrall, 1996).

$$\max \sum_{i=1}^{m} w_i^- s_i^- + \sum_{r=1}^{s} w_r^+ s_r^+$$
$$\text{subject to} \qquad (1.15)$$
$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io} \qquad i = 1,2,...,m;$$
$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{ro} \qquad r = 1,2,...,s;$$
$$\lambda_j, s_i^-, s_r^+ \geq 0$$

where w_i^- and w_r^+ are user-specified weights obtained through value judgment. The DMU_o under evaluation will be termed efficient *if and only if* the optimal value to (1.9) is equal to zero. Otherwise, the nonzero optimal s_i^{-*} identifies an excess utilization of the *i*th input, and the non-zero optimal s_r^{+*} identifies a deficit in the *r*th output. Thus, the solution of (1.15) yields the information on possible adjustments to individual outputs and inputs of each DMU. Obviously, model (1.15) is useful for setting targets for inefficient DMUs with *a priori* information on the adjustments of outputs and inputs.

One should note that model (1.15) does not necessarily yield results that are different from those obtained from the model (1.14). In particular, it will not change the classification from efficient to inefficient (or vice versa) for any DMU.

Model (1.15) identifies a CRS frontier, and therefore is called CRS slack-based model. Table 1.5 summarizes the slack-based models in terms of the frontier types.

Table 1-3. Slack-based Models

Frontier type	Slack-based DEA Model
CRS	$\max \sum_{i=1}^{m} w_i^- s_i^- + \sum_{r=1}^{s} w_r^+ s_r^+$ subject to $\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io} \qquad i = 1,2,...,m;$ $\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{ro} \qquad r = 1,2,...,s;$ $\lambda_j, s_i^-, s_r^+ \geq 0$
VRS	Add $\sum_{j=1}^{n} \lambda_j = 1$
NIRS	Add $\sum_{j=1}^{n} \lambda_j \leq 1$
NDRS	Add $\sum_{j=1}^{n} \lambda_j \geq 1$

1.5. MEASURE-SPECIFIC DEA MODELS

Although DEA does not need *a priori* information on the underlying functional forms and weights among various input and output measures, it assumes proportional improvements of inputs or outputs. This assumption becomes invalid when a preference structure over the improvement of different inputs (outputs) is present in evaluating (inefficient) DMUs (see Zhu (1996b)). We need models where a particular set of performance measures is given pre-emptive priority to improve.

Let $I \subseteq \{1,2, ..., m\}$ and $O \subseteq \{1,2, ...,s\}$ represent the sets of specific inputs and outputs of interest, respectively. Based upon the envelopment models, we can obtain a set of measure-specific models where only the inputs associated with *I* or the outputs associated with *O* are optimized (see Table 1-4).

The measure-specific models can be used to model uncontrollable inputs and outputs (see Banker and Morey (1986)). The controllable measures are related to set *I* or set *O*.

A DMU is efficient under envelopment models if and only if it is efficient under measure-specific models. i.e., both the measure-specific models and the envelopment models yield the same frontier. However, for inefficient DMUs, envelopment and measure-specific models yield different efficient targets.

Consider Figure 1-1. If the response time input is of interest, then the measure-specific model will yield the efficient target of S1 for inefficient S. If the cost input is of interest, S3 will be the target for S. The envelopment model projects S to S2 by reducing the two inputs proportionally.

Table 1-4. Measure-specific Models

Frontier Type	Input-Oriented	Output-Oriented
CRS	$\min \theta - \varepsilon(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+)$ subject to $\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta x_{io} \quad i \in I;$ $\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io} \quad i \notin I;$ $\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{ro} \quad r = 1,2,...,s;$ $\lambda_j \geq 0 \quad j = 1,2,...,n.$	$\max \phi + \varepsilon(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+)$ subject to $\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io} \quad i = 1,2,...,m;$ $\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = \phi y_{ro} \quad r \in O;$ $\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{ro} \quad r \notin O;$ $\lambda_j \geq 0 \quad j = 1,2,...,n.$
VRS	Add $\sum_{j=1}^{n} \lambda_j = 1$	
NIRS	Add $\sum_{j=1}^{n} \lambda_j \leq 1$	
NDRS	Add $\sum_{j=1}^{n} \lambda_j \geq 1$	
Efficient Target	$\begin{cases} \hat{x}_{io} = \theta^* x_{io} - s_i^{-*} & i \in I \\ \hat{x}_{io} = x_{io} - s_i^{-*} & i \notin I \\ \hat{y}_{ro} = y_{ro} + s_r^{+*} & r = 1,2,...,s \end{cases}$	$\begin{cases} \hat{x}_{io} = x_{io} - s_i^{-*} & i = 1,2,...,m \\ \hat{y}_{ro} = \phi^* y_{ro} + s_r^{+*} & r \in O \\ \hat{y}_{ro} = y_{ro} + s_r^{+*} & r \notin O \end{cases}$

1.6. SOLVING DEA WITH DEAFRONTIER SOFTWARE

One can solve the DEA models discussed previously using the spreadsheets and Excel Solver as described in Zhu (2002). In this section, we will demonstrate how to solve the DEA models using the *DEAFrontier* software supplied with the book.

1.6.1 DEAFrontier Software

DEAFrontier is an Add-In for Microsoft® Excel and uses the Excel Solver. This software requires Excel 97 or later versions.

To install the software the CD-ROM using Windows, you may follow these steps:

Step 1. Insert the CD-ROM into your computer's CD-ROM drive. (If the auto run doe not execute, following the following steps.)

Step 2. Launch Windows Explore.
Step 3. Click Browse to browse the CD and find the file "*Setup.exe*".
Step 4: Run "Setup.exe"

DEAFrontier does not set any limit on the number of units, inputs or outputs. However, please check www.solver.com for problem sizes that various versions of Excel Solver can handle (see Table 1-5).

Table 1-5. Microsoft® Excel Solver Problem Size

Problem Size:	Standard Excel Solver	Premium Solver	Premium Solver Platform
Variables x Constraints	200 x 200	1000 x 8000	2000 x 8000

Source: www.solver.com

To run *DEAFrontier*, the Excel Solver must first be installed, and the Solver parameter dialog box must be displayed at least once in the Excel session. Otherwise, an error may occur when you run the software, as shown in Figure 1-5. (*Please also make sure that the Excel Solver works properly. One can use the file "solvertest.xls" to test whether the Excel Solver works. This test file is also available at www.deafrontier.com/solvertest.xls.*)

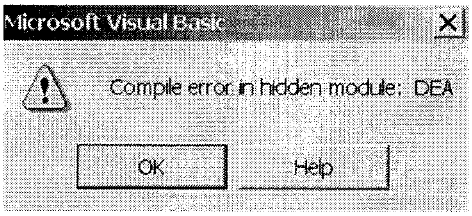

Figure 1-5. Error Message

You may follow the following steps.

First, in Excel, invoke the Solver by using the Tools/Solver menu item as shown in Figure 1-6. This will load the Solver parameter dialog box as shown in Figure 1-7. Then close the Solver parameter dialog box by clicking the Close button. Now, you have successfully loaded the Excel Solver.

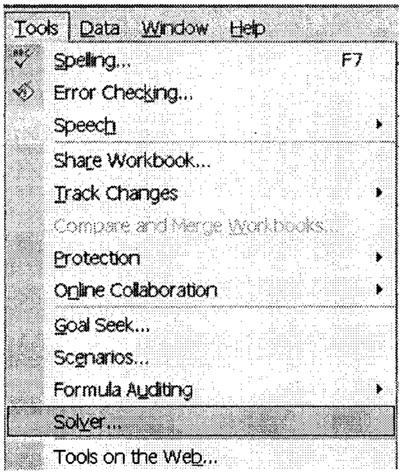

Figure 1-6. Display Solver Parameters Dialog Box

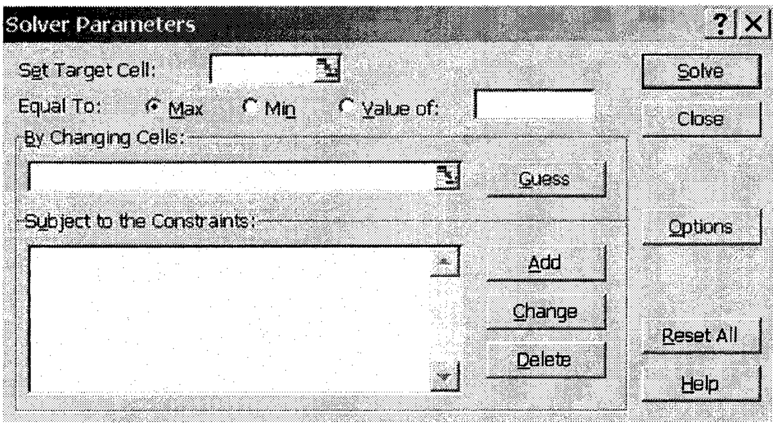

Figure 1-7. Solver Parameters Dialog Box

If Solver does not exist in the Tools menu, you need to select Tools/Add-Ins, and check the Solver box, as shown in Figure 1-8. (If Solver does not show in the Add-Ins, you need to install the Solver first.)

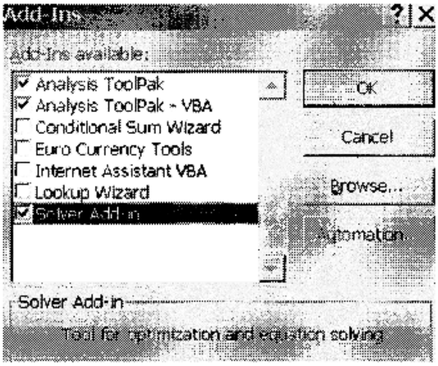

Figure 1-8. Solver Add-In

Next, open the file DEAFrontier.xla, and a "DEAFrontier" menu is added at the end of the Excel menu. (see Figure 1-9). Now, the *DEAFrontier* software is ready to run.

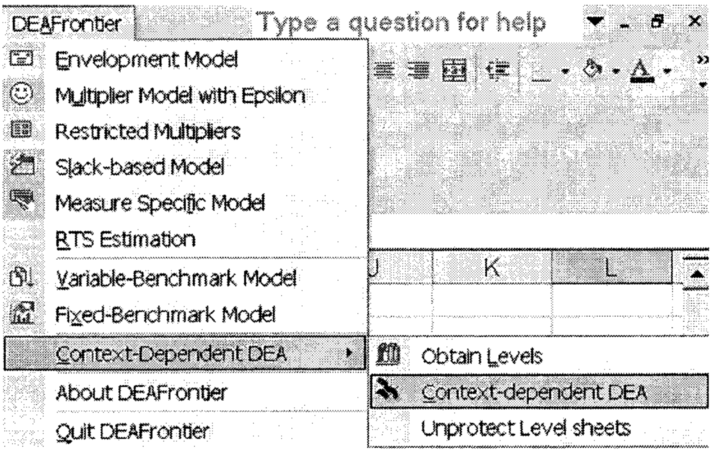

Figure 1-9. DEAFrontier Menu

1.6.2 Organize the Data

The sheet containing the data for DMUs under evaluations must be named as "Data". The data sheet should have the format as shown in Figure 1-10.

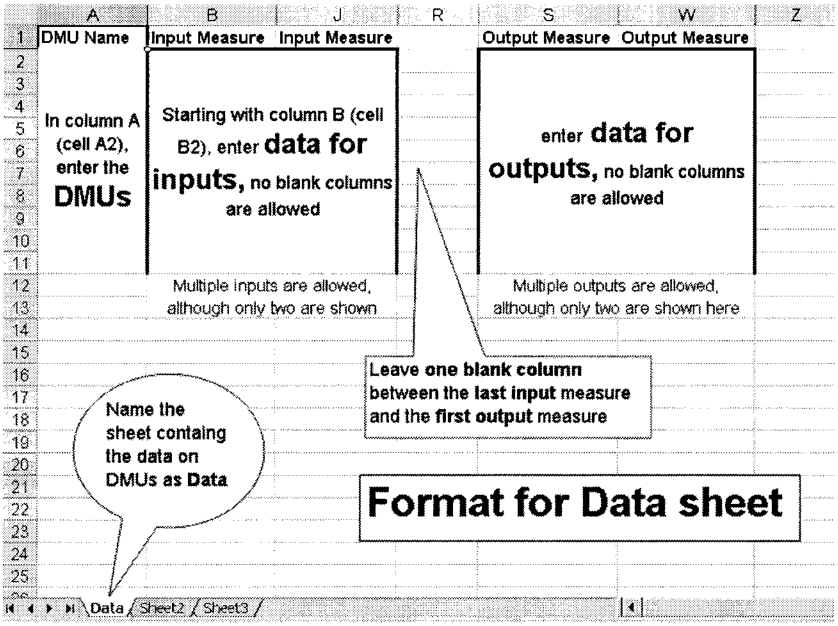

Figure 1-10. Data Sheet Format

Leave one blank column between the input and output data. No blank columns and rows are allowed within the input and output data. Figure 1-11 shows an example where we have top 10 US commercial banks in 1995 with three inputs (employee, assets and equity)and two outputs (market value and profit. (see Seiford and Zhu (1999) for detailed discussion on this data set.)

envelopment model

	A	B	C	D	E	F	G
1	Banks	Employee	Assets	Equity		Market Value	Revenue
2	Citicorp	85300	256853	19581		33221.7	31690
3	BankAmerica Corp.	95288	232446	20222		27148.6	20386
4	NationsBank Corp.	58322	187298	12801		20295.9	16298
5	Chemical Banking Corp.	39078	182926	11912		16971.3	14884
6	J.P. Morgan & Co.	15600	184879	10451		15003.5	13838
7	Chase Manhattan Corp.	33365	121173	9134		12616.4	11336
8	First Chicago NBD Corp.	35328	122002	8450		12351.1	10681
9	First Union Corp.	44536	131880	9043.1		16815	10582.9
10	Banc One Corp.	46900	90454	8197.5		14807.4	8970.9
11	Bankers Trust New York Cor	14000	104000	5000		5252.4	8600
12							

data

Figure 1-11. Sample Data Sheet

Negative or non-numerical data are deemed as invalid data. The software checks if the data are in valid form before the calculation. If the data sheet contains negative or non-numerical data, the software will quit and locate the invalid data.

1.6.3 Run the Envelopment Models

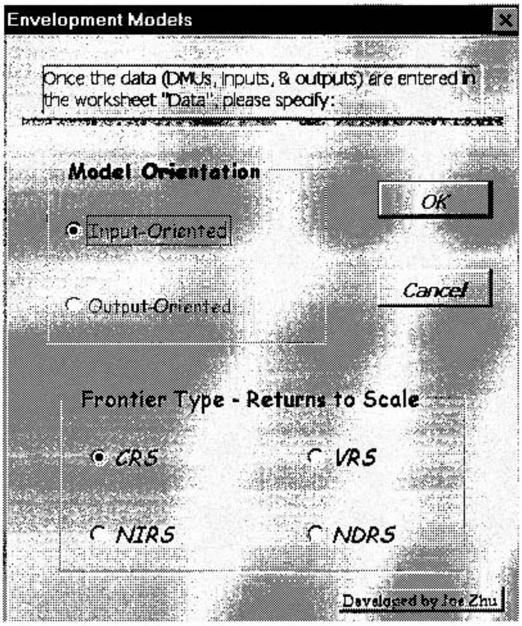

Figure 1-12. Envelopment Model

To run the envelopment models in Table 1-2, select the "Envelopment Model" menu item. You will be prompted with a form for selecting the models, as shown in Figure 1-13.

Model Orientation refers to whether a DEA model is input-oriented or output-oriented and Frontier Type refers to the returns to scale type of the DEA efficient frontier. The software's default selection is an input-oriented CRS model.

The software performs a two-stage DEA calculation. First, the efficiency scores are calculated, and the efficiency scores and benchmarks (Efficiency Reference Set) (λ_j^*) are reported in the "Efficiency" sheet, as shown in Figure 1-14.

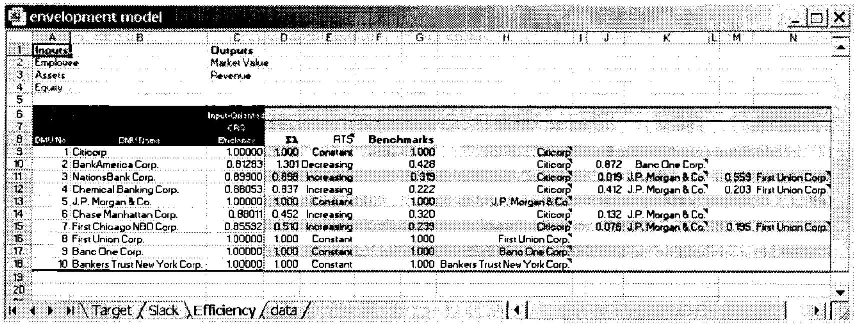

Figure 1-13. Efficiency (Envelopment Model)

The "Efficiency" sheet reports the input and output names. Column A reports the DMU No. Column B reports the DMU names (banks in this case). Column C reports the efficiency scores (it also indicates the type of DEA models used). Column D reports the optimal $\sum\lambda_j^*$ which is used to identify the returns to scale classifications, reported in column E. The Efficiency Reference Set is reported under the "Benchmarks".

At the same time, a "Slack" sheet is generated based upon the efficiency scores and the λ_j^* using the following formula (1.5). Then a "Target" sheet is generated.

Recall that the slacks calculated from (1.5) are not optimized and do not necessary reflect the DEA slack. Therefore, the "Target" sheet may not represent DEA efficient target.

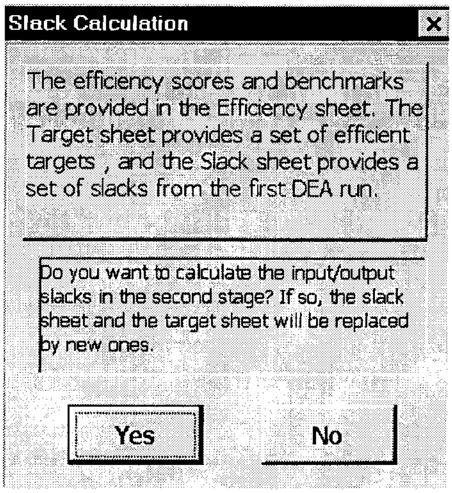

Figure 1-14. Slack Calculation

Therefore, you will be asked whether you want to perform the second-stage calculation, i.e., fixing the efficiency scores and calculating the DEA slacks (see Figure 1-15). If Yes, then the Slack and Target sheets will be replaced by new ones based upon. See file "envelopment model.xls" for the DEA results.

1.6.4 Run the Multiplier Models

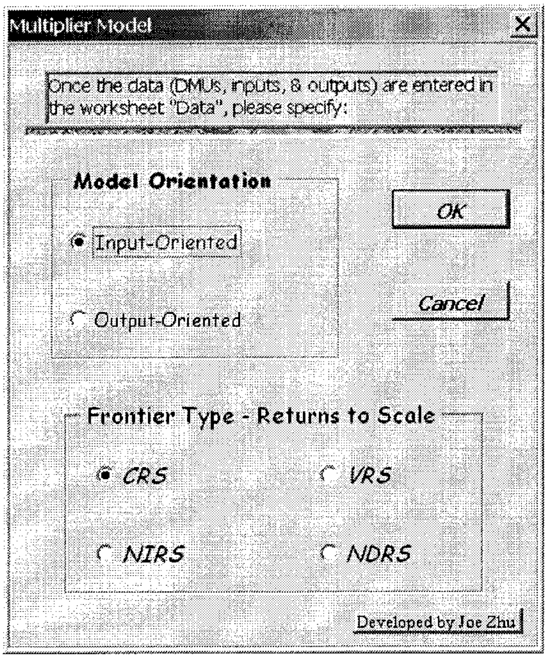

Figure 1-15. Multiplier Model

To run the multiplier models, select the "Multiplier Model" menu item. You will be prompted with a form for selecting the models as shown in Figure 1-15. The form is similar to the one shown in Figure 1-12. The results are reported in a sheet named "Efficiency Report", as shown in Figure 1-16 where the DEA efficiency and optimal multipliers are reported. (Figure 1-16 shows the results of input-oriented VRS multiplier model. See also the file "multiplier model.xls" in the CD.)

multiplier model

	A	B	C	D	E	F	G	H	I
1	**Inputs**		**Outputs**						
2	Employee		Market Value						
3	Assets		Revenue						
4	Equity								
5									
6			Input-Oriented						
7			VRS	Optimal Multipliers					
8	DMU No.	DMU Name	Efficiency	Employee	Assets	Equity	Market Value	Revenue	Free Variable
9	1	Citicorp	1.00000	0.00001	0.00000	0.00001	0.00000	0.00003	0.00000
10	2	BankAmerica Corp.	0.86891	0.00000	0.00000	0.00000	0.00004	0.00000	-0.18650
11	3	NationsBank Corp.	0.90689	0.00000	0.00000	0.00006	0.00003	0.00002	0.07724
12	4	Chemical Banking Corp.	0.90142	0.00000	0.00000	0.00005	0.00003	0.00002	0.13796
13	5	J.P. Morgan & Co.	1.00000	0.00001	0.00000	0.00000	0.00002	0.00003	0.22535
14	6	Chase Manhattan Corp.	1.00000	0.00001	0.00001	0.00000	0.00003	0.00003	0.25951
15	7	First Chicago NBD Corp.	0.94686	0.00001	0.00001	0.00000	0.00003	0.00003	0.25351
16	8	First Union Corp.	1.00000	0.00000	0.00000	0.00009	0.00004	0.00002	0.10780
17	9	Banc One Corp.	1.00000	0.00001	0.00001	0.00000	0.00003	0.00004	0.26544
18	10	Bankers Trust New York Corp.	1.00000	0.00002	0.00001	0.00002	0.00000	0.00012	0.00000
19									
20									
21									

Efficiency Report / data

Figure 1-16. Efficiency Report (Multiplier Model)

1.6.5 Run the AR Models

We need to first set up the sheet "Multiplier" which contains the ARs. For example, if we want to include the following ARs

$$1 \le \frac{v_{Employee}}{v_{Assets}} \le 2.5$$

$$1.5 \le \frac{v_{Employee}}{v_{Equity}} \le 3$$

$$3 \le \frac{\mu_{MarketValue}}{\mu_{\mathrm{Re}venue}} \le 4$$

then the data in the "Multiplier" sheet should be entered as shown in Figure 1-17.

AR model

	A	B	C	D	E
1	1	Employee	Assets	2.5	
2	1.5	Employee	Equity	3	
3	3	Market Value	Revenue	4	
4					

data / multiplier

Figure 1-17. Restrictions on Multipliers

To avoid any errors, we suggest copying and pasting the input and output names from the "data" sheet when entering the information into the "Multiplier" sheet. If the input (output) names in the two sheets do not match, the program will stop.

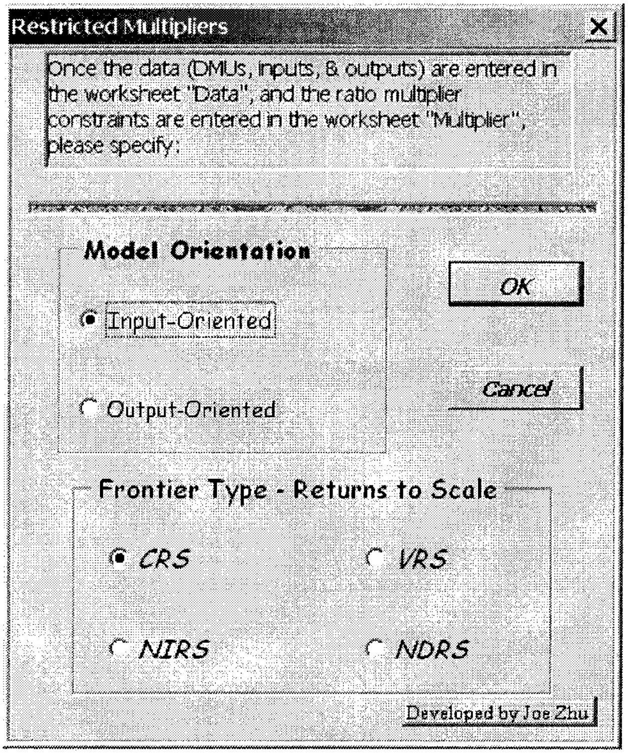

Figure 1-18. AR Model

Once the "Multiplier" sheet is set up, select the "Restricted Multipliers" menu item and you will be prompted to choose a DEA model, as shown in Figure 1-18. Figure 1-19 shows the results of the input-oriented CRS multiplier model with the above ARs.

Note that you can also add ARs that link the input and output multipliers for the "Restricted Multipliers". Note also that if the ARs are not properly specified, then the related DEA model may be infeasible. If that happens, the program will return a value "-9999" for the efficiency score.

Inputs	Outputs
Employee	Market Value
Assets	Revenue
Equity	

DMU No	DMU Name	Input-Oriented CRS Efficiency	Optimal Multipliers Employee	Assets	Equity	Market Value	Revenue
1	Citicorp	1.00000	0.00000	0.00000	0.00000	0.00002	0.00001
2	BankAmerica Corp.	0.80538	0.00000	0.00000	0.00000	0.00002	0.00001
3	NationsBank Corp.	0.84015	0.00001	0.00000	0.00000	0.00003	0.00001
4	Chemical Banking Corp.	0.84197	0.00001	0.00000	0.00000	0.00004	0.00001
5	J.P. Morgan & Co.	0.93846	0.00001	0.00000	0.00000	0.00005	0.00001
6	Chase Manhattan Corp.	0.86080	0.00001	0.00000	0.00000	0.00006	0.00001
7	First Chicago NBD Corp.	0.81792	0.00001	0.00000	0.00001	0.00005	0.00001
8	First Union Corp.	0.92079	0.00001	0.00000	0.00000	0.00005	0.00001
9	Banc One Corp.	1.00000	0.00001	0.00001	0.00000	0.00006	0.00001
10	Bankers Trust New York Corp.	0.63273	0.00002	0.00001	0.00001	0.00008	0.00003

Figure 1-19. AR Results

1.6.6 Run the Slack-based Models

To run the slack-based models, select the "Slack-based Model" menu item. You will be prompted with a form for selecting the models presented in Table 1-3, as shown in Figure 1-20.

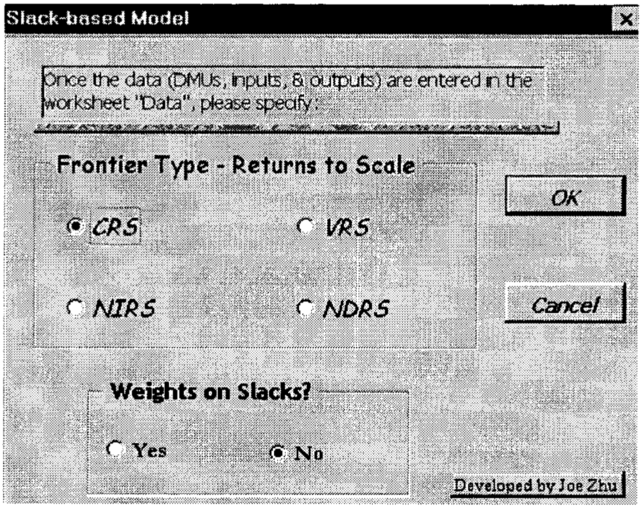

Figure 1-20. Slack-based Models

If you select "Yes" under the "Weights on Slacks", you will be asked to provide the weights, as shown in Figure 1-21. If you select "No", then all the weights are set equal to one.

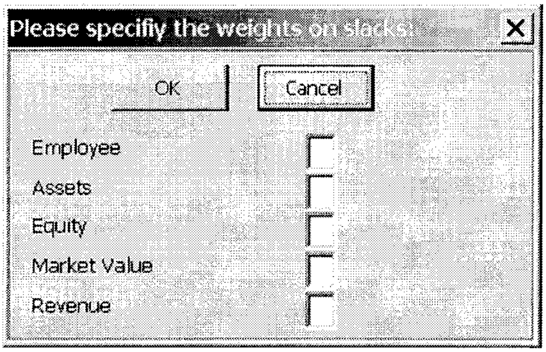

Figure 1-21. Slack Weights

The results are reported in a sheet named "Slack Report" along with a sheet named "Efficient Target". See file "slack model.xls" in the CD.

1.6.7 Run the Measure-Specific Models

To run the measure-specific models, select the "Measure Specific Model" menu item. You will be prompted with a form for selecting the models presented in Table 1-4, as shown in Figure 1-22.

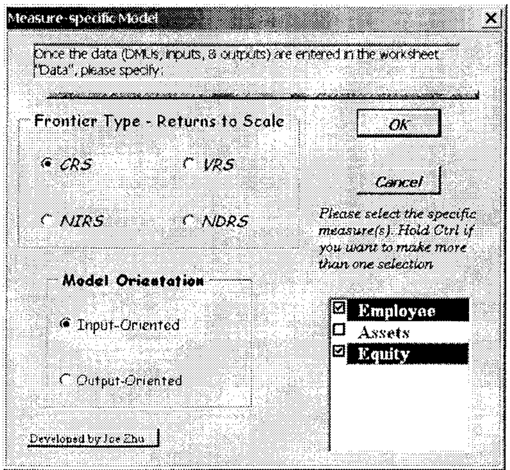

Figure 1-22. Measure-specific Models

Select the measures that are of interest. If you select all the input or all the output measures, then you have the envelopment models.

The results are reported in the "Efficiency", "Slack" and "Target" sheets. See file "measure specific model.xls" in the CD.

REFERENCES

1. Ali, A. I., C.S. Lerme and L.M. Seiford (1995), Components of efficiency evaluation in data envelopment analysis, *European Journal of Operational Research*, 80, 462-473.
2. Banker, ,R.D. and R.C. Morey (1986), Efficiency analysis for exogenously fixed inputs and outputs, *Operations Research*, 34, 513-521.
3. Charnes, A. and W.W. Cooper (1962), Programming with linear fractional functionals, *Naval Research Logistics Quarterly* 9, 181-185.
4. Charnes, A., W.W. Cooper, and E. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2, 429-444.
5. Charnes, A., W.W. Cooper, B. Golany, L.M. Seiford and J. Stutz, (1985), Foundations of Data Envelopment Analysis for Pareto-Koopmans Efficient Empirical Production Functions, *Journal of Econometrics* (1985), 30, pp. 91-l07.
6. Charnes, A., W.W. Cooper, D.B. Sun, and Z.M. Huang (1990), Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks, *Journal of econometrics* 46, 73-91.
7. Cooper, W.W., L.M. Seiford, and K. Tone (2000), Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software, Kluwer Academic Publishers, Boston.
8. Cooper, W.W., L.M. Seiford and J. Zhu (2004), Handbook on Data Envelopment Analysis. Kluwer Academic Publishers, Boston.
9. Dyson, R.G. and E. Thanassoulis, 1988, Reducing weight flexibility in data envelopment analysis, *Journal of the Operational Research Society* 39, No. 6, 563-576.
10. Roll, Y., W.D. Cook, and B. Golany (1991), Controlling factor weights in data envelopment analysis, *IIE Transactions*, 23, 2-9.
11. Seiford, L.M. and J. Zhu (1998), Identifying excesses and deficits in Chinese industrial productivity (1953-1990): A weighted data envelopment analysis approach. *OMEGA, International Journal of Management Science* 26, No. 2, 279-296.

12. Seiford, L.M. and J. Zhu (1999), Profitability and marketability of the top 55 US commercial banks, *Management Science*, Vol. 45, 1270-1288.
13. Thompson, R.G., F.D.Jr. Singleton, R.M. Thrall, and B.A. Smith (1986), Comparative site evaluation for locating a high-energy physics lab in Texas, *Interfaces* 16, 35-49.
14. Thompson, R.G., L. Langemeier, C. Lee, E. Lee, and R. Thrall (1990), The role of multiplier bounds in efficiency analysis with application to Kansas farming, *Journal of Econometrics* 46, 93-108.
15. Thrall, R.M. (1996), Duality, classification and slacks in DEA, *Annals of Operations Research,* 66, 109-138.
16. Wong, Y.-H.B. and J.E. Beasley (1990), Restricting weight flexibility in data envelopment analysis, *Journal of the Operational Research Society* 41, 829-835.
17. Zhu, J. (1996a), DEA/AR analysis of the 1988-1989 performance of the Nanjing Textiles Corporation, *Annals of Operations Research* 66, 311-335.
18. Zhu, J. (1996b), Data envelopment analysis with preference structure, Journal of the Operational Research Society 47, No. 1, 136-150.
19. Zhu, J. (2002), Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets and DEA Excel Solver, Kluwer Academic Publishers, Boston.

Part of this chapter is based upon chapter 1 in Zhu, J. (2002), Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets and DEA Excel Solver, Kluwer Academic Publishers, Boston

Chapter 2

MEASURING EFFICIENCY OF HIGHWAY MAINTENANCE PATROLS

2.1. BACKGROUND

A number of applications of DEA are found in the area of maintenance. In the particular application discussed in this chapter, we look at the performance of highway maintenance crews or patrols in the province of Ontario, Canada. The discussion herein is based on the work of Cook et al (1990), (1991), and (2001). The problem of measuring efficiency in the roadway maintenance sector is an important one that has been examined by others as well. Deller and Nelson (1991), for example, examined a similar problem but where network size is used as an output and material, labour and capital are inputs. Later, Rouse et al. (1997) revisited the road maintenance problem for the case of highways in New Zealand, by considering additional inputs and outputs. In particular, they attempt to address environmental differences among patrols by incorporating factors aimed to capture geological indicators. This was undertaken, presumably in realization of the fact that patrols are not necessarily comparable via the conventional inputs and outputs. As well, they attempt to pay attention to weight restrictions as raised earlier by Roll, Cook and Golany (1991).

At the time that the initial study of Cook et al. (1990) was conducted, the stipulated rationale for having a formal performance measure for each patrol was to permit budget setting in a resource constrained environment. As funding for maintenance has eroded over time, a need has arisen for a formal mechanism whereby patrols are treated equitably in regard to the allocation of maintenance dollars. What is most appealing about the DEA rationale in this setting, is that if an inefficient patrol can attain efficiency status, its projected inputs θX_o can aid in setting its budget, where X_o is the vector of Inputs.

Most of the routine maintenance activities on Ontario's highways fall under the responsibility of the 244 patrols scattered throughout the province. Each such patrol is responsible for some fixed number of lane-kilometers of highway, and those activities associated with that portion of the network. More than 100 different categories of operations or activities exist, and are grouped under the headings: 'surface,' 'shoulder,' 'right of way,' 'median,' and 'winter operations.'

The present system for monitoring patrol activities is the Maintenance Management System (MMS). This is a computerized record keeping system which keeps track of total work accomplished by type of operation, patrol and highway class. This system is similar to those in other Canadian provinces and states in the U.S.A.

While various statistics (such as median operations accomplished, by highway class) are maintained, there is presently no formal process for evaluating patrol activities. An area of importance to the Ministry has to do with the efficiency with which maintenance operations are carried out in various parts of the province. Since observed accomplishments influence budgetary decisions, a better understanding of efficiency will give management a yardstick for measuring what accomplishments can be expected within a given budget limit.

While there are various possible approaches to the problem of measuring efficiency in this context, the DEA framework is particularly appropriate for a number of reasons. First, the prospect of obtaining "production standards" in the usual engineering sense seems doubtful. The number of different "products" and different environmental and soil conditions mitigate against a conventional industrial engineering approach. Second, DEA is capable of handling non-economic factors, like number of accidents, maintenance dollars (an economic factor), cars/day, average age of pavement, etc., and allows for measurement of such factors on different scales. Such an approach seems particularly suited to the maintenance area, since factors such as traffic intensity, safety parameters and average age of pavements are an important part of the picture.

These and other reasons point to the appropriateness of DEA.

2.2. DEA ANALYSIS

2.2.1 The Model and its Factors

In a study of potential factors which could be utilized to best represent causes and effects relating to patrol performance, four outputs and three inputs have been chosen. Specifically, the efficiency e is given by

$$e = \frac{u_1(ASF) + u_2(ATS) + u_3(RCF) + u_4(APF)}{\upsilon_1(MEX) + \upsilon_2(CEX) + \upsilon_3(CLF)},$$

and (u_1, u_2, u_3, u_4) and $(\upsilon_1, \upsilon_2, \upsilon_3)$ denote output and input factor weights respectively.

ASF – Area Served Factor

This factor was chosen to measure the *extent* of the work load for which the patrol has responsibility. The ASF factor value is calculated from the formula

$$\text{ASF} = \sum_i \left[L_i (TLE)_i \left(A_j + C\right) + L_i \left(S_i B_j + D\right)\right]$$

where:

L_i – Length of road section *i*
TLE_i – Two-lane equivalent of road section *i*
S_i – Shoulder width of road section *i*
A_j – Coefficient for road surface type *j* (the one in road section *i*)
B_j – Coefficient for shoulder type *j* (the one in road section *i*)
C – Coefficient for winter operations
D – Coefficient for other operations (ROW, median etc.)

ATS – Average Traffic Served

This factor is intended to be a measure of the overall benefit to the users of the highway system in a patrol. The formula for computing ATS is given by

$$\text{ATS} = 10^{-3} \sum_i L_i \text{ (AADT)}_\text{I},$$

where AADT_i is the Annual Average Daily Traffic and 10^{-3} is a scaling factor designed to bring ATS within a reasonable range for analysis.

RCF – Pavement Rating Change Factor

This factor measures the actual change in PCR, (Pavement Condition Rating) of the various road sections, relative to a 'standard' change for the same period.

APF – Accident Prevention Factor

Much of the work of maintenance staff arises due to the need to prevent accidents (surface & shoulder repairs, washouts, etc.) In this regard, accident prevention can be viewed as a cause or goal of maintenance.

A reasonable measure of accident prevention should be directly proportional to traffic level (ATS), and inversely proportional to the observed number of accidents. The chosen form is given by

$$\text{APF} = 100\frac{ATS}{C},$$

where 100 is a scaling factor and C is the number of road accidents, during the observed period, on all road sections serviced by a patrol.

MEX – Maintenance Expenditures

This is the total of all expenditures linked to the patrol. It includes both "in-house" work as well as maintenance activities performed by private contractors. Moreover, MEX includes any district-supplied services such as equipment and district supervisors' salaries.

CEX – Capital expenditures

This is the total of all capital expenditures made toward improving the existing highway infrastructure. This would include resurfacing, shoulder paving, repairs to structures, dome construction, etc. – all activities which complement maintenance efforts. Excluded are new link and new structure construction, since these do not directly complement maintenance.

CLF – Climatic Factor

What can often be an overriding consideration in the performance of a patrol, is the environmental circumstances in which that patrol must operate. The amount of snowfall, for example, will clearly influence the level of winter maintenance (snow removal and salting) needed. The extent of spring breakups will directly influence the need for summer road surface work.

Four sub-factors were taken into account in arriving at an overall climatic factor:

Snowfall
Major temperature cycles
Minor temperature cycles
Rainfall

Available data from weather stations were used to compute these sub-factors.

The overall climatic factor for a patrol is computed from:

$$\mathrm{CLF_k} = \sum_i P_{ki}(\sum_j (W_j / D_{ij})),$$

where

k – patrol index;

P_{ki} – weight of station i in calculating the climatic factor of patrol k

W_j relative importance weight of climatic factor j.

$W_1 = 50$
$W_2 = 300$
$W_3 = 20{,}000$
$W_4 = 1{,}000$

It is noted that the weights W_j were chosen while taking into account the numerical scales of each of the climatic factors (e.g. the snowfall numbers are much greater in size than the major cycle numbers). In addition, the weights were selected with attention to the resultant CLF measure being relatively of the same order of magnitude as the other efficiency factors.

2.2.2 Data and Unbounded Runs

In the present study, 4 districts are used, having a combined total of 62 patrols. As an illustration, the factor values for one of the patrols are given by:

ASF = 404
ATS = 267
RCF = 184
APF = 331
MEX = 585
CEX = 284
CLF = 715

The first level of analyses carried out uses the entire set of patrols, with 62 L.P. problems being solved. It is noted that the only constraints other than the ratio restrictions (converted to linear format) are restraints stipulating that all variables should be nonzero. This means that no patrol is permitted to assign an importance of 0 to any factor. The model is therefore, referred to as the 'unbounded' model. (The bounded model, to be discussed, will contain significant upper and lower bounds on the variables).

The results from the 62 unbounded runs are shown under column (1) of summary Table 2-1. Note the rating of 0.725 for the first patrol in District 2.

Table 2-1. Summary of Efficiencies

DMU		Efficiencies			
		Unbounded	CSW	Indiv. Wgts Entire Sample	Indiv. Wgts Within District
D	P	1	2	3	4
2	1	.725	.517	.655	.724
	3	.768	.654	.741	.878
	5	.663	.540	.614	.795
	6	.700	.606	.671	.791
	7	.650	.545	.622	.725
	9	.739	.581	.679	.751
	10	.841	.675	.756	1
	11	.948	.699	.836	1
	13	.951	.786	.912	1
	15	1	1	1	1
	16	1	.569	.761	.776
	17	1	.664	.891	.931
	18	.857	.704	.774	.839
3	1	.855	.608	.722	.988
	2	.787	.640	.744	1
	3	.756	.492	.642	.793
	4	.761	.658	.752	1
	5	1	.497	.641	.842
	6	.990	.695	.874	1
	7	1	.661	.945	1
	8	.840	.637	.811	.982
	9	.944	.624	.786	.991
	10	.613	.443	.562	.693
	11	.802	.587	.729	.923
	12	1	.528	.705	.941
	13	.921	.520	.677	.881
	14	.457	.380	.430	.608

Table 2-1 continued

DMU		Efficiencies			
		Unbounded	CSW	Indiv. Wgts Entire Sample	Indiv. Wgts Within District
D	P	1	2	3	4
8	1	.867	.528	.672	.246
	2	.885	.181	.353	.353
	3	1	.813	.940	1
	4	.998	.822	.881	.925
	5	1	.966	1	1
	6	1	.757	.849	.912
	7	.876	.745	.824	.906
	8	1	.852	.949	.984
	9	1	.727	.836	.945
	10	1	.706	.980	1
	12	1	.832	.903	.955
	13	1	.673	.869	.939
	14	1	.739	.879	.956
	15	1	.430	1	1
	16	.871	.772	.825	.883
	17	.942	.803	.872	.897
	18	1	.591	.763	.870
	19	.826	.648	.753	.885
	21	1	.724	.848	.953
	22	1	.745	1	1
	25	.817	.659	.799	.875
20	1	.583	.440	.526	.526
	2	.739	.213	.369	.370
	3	.989	.343	.541	.541
	4	1	.746	.914	.974
	5	.770	.593	.674	.674
	6	1	.836	.948	.966
	7	1	.964	1	1
	8	.781	.455	.597	.600
	9	.915	.449	.625	.635
	10	1	.710	.890	.921
	11	.819	.419	.586	.586
	12	.927	.609	.762	.768
	13	.933	.670	.795	.843
	14	1	.643	.849	.849

2.2.3 Bounded Runs

It must be emphasized again that the unbounded model yields efficiency ratings that tend to credit the patrol with a higher level of performance than may be justified. Since complete flexibility in choice of weights is permitted, the model will often assign unreasonably low or unreasonably high weights (multipliers) to some factors in the process of trying to drive the efficiency rating for the patrol in question as high as possible. Moreover, the weight assigned to a factor (e.g. CEX) by one patrol may differ drastically from the weight assigned to that factor by another patrol. Thus, in order to exercise some reasonable level of control over the manner in which importance weights are assigned, bounds need to be imposed in the model.

Given a set of absolute bounds L_i^1, U_i^1 on output multipliers and L_j^2, U_j^2 on inputs, the constraints $L_i^1 \leq \mu_i \leq U_i^1$ and $L_j^2 \leq \upsilon_j \leq U_j^2$ are added to model (1.3) of Chapter 1.

The efficiency ratings resulting from runs of this bounded version of the model are displayed in column 3 of Table 2-1. It is noted that the efficiencies obtained from the bounded runs are lower than or equal to the corresponding efficiencies arising from the unbounded analysis.

2.2.4 Deriving a Common Set of Weights

A case can be made, however, for having a Common Set of Weights (CSW). Being able to evaluate all patrols from a common reference point provides one basis for rank ordering the DMUs from best to worst. While no "best" method exists for determining such a set of weights, a simple procedure was developed for the organization in question.

Briefly, the procedure works as follows: Choose the highest priority factor, (e.g., μ_1), and while restricting all factor weights to be within their respective bounds, maximize (or minimize) the weight for the factor in question. In this particular case μ_1 is chosen as a first priority since it is both a reliable measure of output and is believed to strongly affect efficiency. The factor weight is maximized if the indicated direction is "up," and is minimized if the direction is "down."

When the optimal weight value (e.g. $\mu_1 = 800$) is determined, it is then fixed at that level in the later optimization stages. The next factor in priority is then chosen (e.g. υ_3), and minimized subject to the same constraints as applied previously, but with $\mu_1 = 800$. This process is continued until all factor weights have been set and the Common Set of Weights is established.

Efficiency ratings using the CSW are shown in column 2 of Table 2-1. Note that patrol 15 in District 2 has an efficiency rating of 1.0, when using these weights. Thus, at least in this case, the CSW is feasible.

2.2.5 District Runs

In order to extract maximum information for effective managerial control, the DEA model was run for each district separately. The resultant set of district efficiencies appears in Column 4 of Table 2-1. It is noted that these district efficiencies are higher than the corresponding values obtained when the entire set of patrols was considered. The smaller comparison groups in the district analyses give rise to this phenomenon. It is also the case that some patrols which were inefficient in the earlier analysis, obtained a rating of 1.0 in the district setting, since those efficient patrols in *other* districts against which comparison was made have been removed from the peer group.

Because significant differences may exist from one district to another (for example, climatic and highway type differences), the intra-district efficiency measures of column 4 in Table 2-1 may provide a fairer appraisal of performance. At the same time, it is desirable to detect any district-to-district differences, necessitating inter-district comparisons. Overall district performance can be viewed in a number of ways. Two useful measures that can be derived are technical efficiency and managerial efficiency.

Technical Efficiency – with this measure we compare "best" performance in a district to best performance in another district. This is taken as an indicator of the 'technical potential' of a district. Simply speaking, technical efficiency is a measure of the distance of the district frontier from the overall frontier.

One technique for obtaining this measure is to bring all points in a district to the district frontier by applying the "adjustment" method proposed in Charnes et al (1978). A somewhat simpler approach is to "correct" the district efficiencies by dividing the overall efficiency of each patrol (column 3 of Table 2-1) by the relative efficiency within the district (column 4). The resulting quotients are approximations of individual patrol efficiencies if they were brought to the district frontiers.

Taking the average of all corrected efficiencies within a district is then a measure of technical efficiency. These values are shown in column 2 of Table 2-2. it is noted, for example, that the best performance of district 20 (.986) is near the best for the entire group. District 3 on the other hand has its best performers only at 79% of the overall best performance.

Managerial Efficiency – this measure refers to the actual performance of patrols, rather than that of best performers as above. The most reasonable measure to take is the average of the actual efficiencies for the patrols in a district. Column 1 in Table 2-2 provides the average of efficiencies when the comparison group is the overall set. Column 3 is the average when the comparison group is only that set of patrols within the district. Naturally, the latter average (column 3) is larger than the former (column 1).

Table 2-2. District Efficiencies

District	# of patrols	(1) Eff. of relative to overall frontier	(2) Ave. eff. dist. front. relative to over. front.	(3) relative to district frontier
2	13	.762	.884	.862
3	14	.716	.790	.903
8	21	.847	.938	.904
20	14	.720	.986	.732

It is noted that the managerial efficiency relative to the entire group is *approximately* equal to the product of the managerial efficiency relative to the district and the technical efficiency of the district. Exact equality fails here because of the manner in which the averages are obtained.

2.2.6 Analysis of Various Characteristics

Over and above the input parameters chosen for the analysis of patrols, there are other influences (on performance) which deserve attention. These influences can be thought of as *characteristics* or circumstances which can affect the efficiency with which a patrol operates. Two particular characteristics have been chosen:

(1) % privatization
(2) traffic level

The method used to examine a given characteristic was to (1) define levels for that characteristic, (2) separate out those patrols corresponding to the various levels, and (3) do a separate analysis on each of the subgroups arising from this separation process. As an illustration, consider % privatization. Here, a particular level (for example, 10%) was chosen as the threshold separating "high" from "low" privatization. Those patrols with a percentage at or below 10% were then subject to the aforementioned analyses. This was then repeated for patrols above 10%.

The percentage of privatization is defined as the proportion of the total maintenance budget for the patrol which is utilized on privatized jobs. The proportion can be determined from the budget codes provided in the data file from which the financial information was extracted. As an example of the type of analysis which would proceed from the setting of a threshold level, the following displays the results for District 8. (See Table 2-3).

Table 2-3. Analysis by % Privatization: District 8

	Subgroup A (above 10%)		Subgroup B (below 10%)	
Patrol #	D	A	D	B
1	.7458	.7730		
2	.3530	.4560		
3			1	1
4	.9248	1		
5			1	1
6	.9120	.9255		
7			.9060	.9060
8	.9835	1		
9			.9450	.9450
10			1	1
12			.9553	.9553
13			.9384	.9387
14			.9561	.9679
15			1	1
16			.8829	.8988
17			.8975	.8992
18	.8701	.8733		
19	.8847	.8893		
21	.9529	.9789		
22	1	1		
25			.8747	1
$\sum$	7.6268	7.8360	11.3559	11.5109
Av.	.8474	.8773	.9463	.9592

Table 2-4. District 8. Sub-group A: above 10%. Sub-group B: below 10%

		Average efficiencies	
	Number of DMUs	District Analysis	Sub-group analysis
Sub-group A (high privatization)	9	.8474	.8773
Sub-group B (low privatization)	12	.9463	.9592
Total/Average`	21	.9039	

The column labeled "D" provides the overall district efficiencies which were presented earlier and have been obtained without consideration of privatization influences. When those patrols in district 8 with privatization below 10% are examined separate from the rest, different efficiency ratings result. These are displayed under column A. Note, for example, that the rating for patrol 1 rises from .7458 to .7730. Recall that the rating for a patrol when looked at in the presence of a subgroup will always be at least as high as is the corresponding "entire group" rating. The results of this type of analysis can be summarized in terms of averages, as per Table 2-4.

As a general rule, when looking at changes in average performance from the "entire district" results to the subgroup (say low privatization) results, *small* changes point to a *positive* influence of the level of the characteristic corresponding to that subgroup. For example, in the case of low privatization in district 8, the average efficiency rating of .9592 is not significantly different than the average for these patrols when analyzed relative to the entire district (.9463). This can only be explained by the fact that very few high privatization patrols were on the frontier. Thus, low privatization patrols tend to perform better than high privatization patrols since more of the former were on the frontier than was true of the latter. On the other hand, the average efficiency rating for high privatization patrols jumped from .8474 to .8773. This means that some improvement in the performance picture for high privatization patrols occurs when the efficient low privatization patrols are removed from the analysis.

As to possible inferences which one might make in the case of, say, district 8, patrols practicing a low privatization policy tend to perform on average better than is true of those with high privatization. In the case of patrol #2, for example, 0.103 points out of the total efficiency gap of .647 (=1 - .353) can be explained by privatizing out a large proportion ($\approx$ 11%) of its work.

In general, privatization impacts are different from district to district. Overall there is no conclusive evidence that privatization increases efficiency. In fact the converse seems to be true in the case of district 20.

2.3. OUTPUT DETERIORATION WITH INPUT REDUCTION

2.3.1 Theoretical versus Achievable Targets

As with many applications of DEA, implementation in the maintenance crew setting has revealed a gap between the theoretical and realistically achievable resource reduction in inefficient units. Specifically, for a given inefficient patrol, the actual input reduction $(1-\alpha)$ deemed feasible by the maintenance supervisor and geotechnical staff, who have intimate knowledge of that patrol's highway network, generally falls short of the DEA-derived $1\text{-}\theta$ for that DMU. There is a belief that below the αX_0 level, the remaining resources would not be sufficient to keep the roadway at the same standard as is currently experienced by that DMU. The general explanation for this is that the frontier units that act as peers for such inefficient units, may be operating in a more favorable environment. In the highway setting, this can mean that the frontier units may be achieving efficiency partially because highway surface conditions are superior to those of inefficient units, or that roadway sub-grade structures result in slower deterioration in the peer patrols. As well, the model of Cook et al. (1990) fails to account for certain environmental factors such as average daily temperature.

Some attempt was made in the earlier study to control for road condition, by way of a non-discretionary input, *the average pavement rating*. This rating is, however, generally not adequate to reflect the level of ongoing maintenance needed to maintain a certain standard. This rating primarily captures visible surface conditions such as extent of pavement cracking, number and severity of ruts and potholes, etc. It would not account for sub-grade depth, total pavement thickness and so on. If kept at a desirable standard, the roadway would be expected to achieve a certain life expectancy before major rehabilitation is required. If available resources are reduced below some critical point αX_o, however, a faster deterioration would result, and the expected useful lives of roads in that patrol would be reduced.

In an attempt to provide a more acceptable DEA methodology (that would be accepted by management within the transportation ministry), the earlier model of Cook et al. (1990) was upgraded to include a provision for climatic conditions. This was done in recognition of the fact that severity of snowfall clearly influences winter maintenance expenses, while the amount of rainfall impacts summer maintenance. Cook et al (1994) present an upgraded version of the earlier model that incorporates these factors, as well as a delineation between summer and winter traffic conditions. Even with this further allowance for environmental differences, however, many patrols are still unable to achieve computed performance targets, and argue that significant anomalies still exist.

Rouse et al. (1997) experienced a similar problem, and introduced a categorical variable in an attempt to address environmental differences that exist among patrols. As presented by Banker and Morey (1986), categorical variables are intended to recognize different environments in which DMUs may operate. See also Rousseau and Semple (1993). Essentially, if the setting is one where there is a single dimension (e.g. size of bank branch) according to which DMUs can be grouped, so that those in the same category are clearly comparable, then this enhanced model structure might solve the aforementioned problem of DMU anomalies. In an attempt to apply this logic in the maintenance patrol setting, however, the authors found that there was no such *single* dimension along which patrols could be ranked. For example, much of the winter and spring maintenance is a function of snowfall, temperature, temperature fluctuations, number of freeze/thaw cycles, etc. Patrols in the north do experience lower winter temperatures, thus causing pavements there to break up more rapidly than is true in *similar* patrols with more favorable temperatures. Thus, one might be tempted to categorize patrols according to temperature (or even total days of extreme cold weather). Unfortunately, it is the number of *freeze/thaw cycles* which can cause even more pavement surface damage (although geotechnical research fails to capture precisely how much more damage). It turns out to be the case that northern patrols suffer fewer such cycles than is true of patrols in more favorable temperate locations (i.e. southern patrols). One could also point to non-climate related factors, such as extent to which sub-grades under road surfaces are influenced by poor drainage conditions (e.g. swampland). A factor such as this might serve as a categorical variable as well.

The conclusion of this investigation was that categories of DMUs could be formed in several (often conflicting) ways. While it is true that more than a single categorical input can be included, meaning that a *partial* ordering of the data is possible (see e.g., Cooper, Seiford and Tone (2000)), in the present circumstances there appeared to be so many different dimensions on which DMUs could be categorized, that the model became somewhat indeterminate. This fact rendered the categorical variable approach rather inapplicable in the environment examined.

2.3.2 Enforced Input Reduction

The conventional application of DEA (for example, the VRS input-oriented model of Banker et al. (1984)), may not be appropriate in many settings for at least two reasons. First, the projection to the frontier may not be 'slackless', which will occur if a DMU is improperly enveloped. Thus, the very idea that in order to reach a projection *on* the frontier, outputs may actually have to increase, for example, renders the model rather unrealistic in a setting where the outputs are *traffic served* and *area.* Arguably, increased outputs here can mean performing a level of maintenance above that which is currently the practice, hence providing a better and more serviceable roadway for those drivers who do use it. The second, and more serious restriction of the DEA structure, is that even if one acknowledges that a radial reduction in inputs by a factor $1-\theta$ is not feasible, there is the common presumption that a reduction of a lesser amount $1-\alpha$ (where $\alpha > \theta$) will be acceptable to management. The problem here is that even if it is accepted that a given patrol cannot forfeit more resources than $(1-\alpha)X_o$, and still provide the same level of service, budget realities can deem it *necessary* to operate with *less* resources than this level dictates. Thus, budgetary reality calls for *enforced input reduction*, often beyond the αX_o critical level. Such enforced reduction in inputs is generally accompanied by *erosion of outputs*.

The important feature of the efficiency measurement exercise here, is that the measure itself is simply a means to an end. Management wishes to use such measures as a mechanism for establishing an appropriate level of maintenance funding within the province. Equally important, it wishes to gauge the impact on the highway system in the common event of under-funding. What will be the extent of the damage to the serviceability of the highway? What are the long run implications of reduced maintenance on future capital reconstruction of the highway network?

In the event where less resources are available than needed to meet standards, management's course of action would depend on the problem setting. In a bank branch situation, for example, inadequate resources, (for example branch personnel), might simply mean that there will be longer waiting times for customers, more complaints, lost accounts, and reduced sales of financial services products. In the long run, performance suffers through deteriorating sales, and overall transactions; that is, outputs decline. In the maintenance setting, inadequate resources could result in some maintenance activities being uniformly discontinued throughout the patrol area (e.g., crack sealing could be halted, roadside activities such as grass cutting might be done less often, etc.). Alternatively, management may choose to maintain the higher traffic-volume roads to standard, while sacrificing maintenance work on less important ones. Thus, on average, the serviceability, hence the output deteriorates.

The principle issue that maintenance management now faces is to obtain not only a measure of the theoretical efficiency vis-a-vis a frontier of best performing patrols, but, as well, to evaluate this against practically achievable targets. At the same time, as indicated above, management wants to assess the likely decline in roadway standards, should an inefficient patrol be required to achieve *frontier status*. Such information can aid management in setting budget targets. Specifically, reduced standards in a patrol can have long term implications for drivers (in the form of rougher roads), and for the government agency, and ultimately the taxpayer, in the form of more frequent capital expenditures prompted by shortened pavement lives. Savings in present day maintenance expenditures would, therefore, need to be traded off against accelerated resurfacing and reconstruction options.

2.3.3 Modeling Output Erosion

Let us now examine the phenomenon of output decline within the DEA context. Assume that there are n decision making units, R outputs and I

inputs, and consider the variable returns to scale (VRS) model of Banker et al. (1984) for development purposes herein. Let X_j, Y_j denote respectively the vectors of inputs and outputs for DMU $_j$. For purposes of exposition, we also assume in this section that all variables are discretionary. In the example of the following section, however, certain variables are nondiscretionary, and are treated as such.

The ratio form of the variable returns to scale model of Banker, Charnes and Cooper (1984) (BCC), is given by:

$$\begin{aligned}
&\max \frac{uY_o + \omega}{vX_o} \\
&\text{subject to:} \\
&\frac{uY_j + \omega}{vX_j} \leq 1, \qquad j = 1,...,n \\
&u, v \geq 0, \quad \omega \quad \text{unrestricted}
\end{aligned} \tag{2.1}$$

The linear programming equivalents (dual and primal problems) are:

$$\begin{aligned}
&\max\ \mu Y_o + \omega \\
&\text{subject to} \\
&vX_o = 1 \\
&\mu Y_j + \omega - vX_j \leq 0, j = 1,...,n \\
&\mu, v \geq 0, \qquad \omega \text{ unrestricted}
\end{aligned} \tag{2.2}$$

and

$$\begin{aligned}
&\min \theta \\
&\text{subject to:} \\
&\theta X_o - \sum_{j=1}^{n} \lambda_j X_j \geq 0 \\
&\sum_{j=1}^{n} \lambda_j Y_j \qquad \leq Y_o \\
&\sum_{j=1}^{n} \lambda_j = 1 \\
&\lambda_j \qquad \geq 0, \ j = 1,...,n
\end{aligned} \tag{2.3}$$

As indicated above, earlier attempts to include environmental variables, and to introduce categorical inputs failed to produce targets which many patrols deemed achievable. Hence, management has tended to *adjust* DEA targets to better reflect the reality existing in certain patrols. Specifically, patrol supervisors, in collaboration with geotechnical engineers, and regional office maintenance managers, have specified what they perceive as the maximum possible input reductions $(1-\alpha_j)\%$ in respective patrols j. These values are set with the understanding that if a reduction of more than $(1-\alpha_j)\%$ in all discretionary inputs (primarily the maintenance budget) should occur in patrol j, it is claimed that outputs will begin to *erode* by some percentage γ_j. Output erosion generally means that a lower *quality* of road maintenance is being administered, as discussed in the previous section. As indicated above, the visible consequence of insufficient resources in a patrol can mean the equivalent of discontinuing maintenance on a portion of the network. To put this in context, note that the outputs we have used in the previous study are *traffic* (total users served), and *area* (roadway and roadside combined) maintained. Reduced outputs can be viewed as fewer road users receiving adequate services.

Let us assume for purposes of model development in this section, that declared expectations of output erosions are provided in good faith and represent reality. Clearly, there can be an incentive for the patrol supervisor to overstate potential output erosion, making intended budget reductions appear highly undesirable from management's perspective. There are a sufficient number of patrol-specific anomalies, such that impacts of budget reductions can only be truly estimated by the maintenance supervisor and accompanying geotechnical staff of that patrol. Hence, senior (head office) management could potentially be 'at the mercy' of patrol staff in regard to honest declarations.

One has to remember, however, that certain realities do make it rather difficult if not impossible, for patrol management to cheat in this regard. First, geotechnical staff is generally shared by several patrols, meaning that there would be little incentive to exaggerate the resource needs of one patrol at the expense of another. As well, the claims of one district supervisor must hold up to scrutiny by other supervisors who compete for the same resources. The modeling considerations discussed herein are, therefore, correct and relevant only to the extent that erosion rates reflect what will actually happen. Issues pertaining to obtaining accurate estimates of output deterioration in patrols are, thus, primarily behavioral in nature, and beyond the scope of this research.

To model the output deterioration phenomenon, refer to Figure 2-1. Note that in this simplified image of projection, with a single input and single output, inputs are reduced with no impact on outputs up to the point $\alpha_o X_o$. From that point on, outputs are assumed to radially deteriorate at a rate of γ_o

per radial percentage unit reduction in X_o, finally projecting to a level $\hat{Y}_o$ on the frontier.

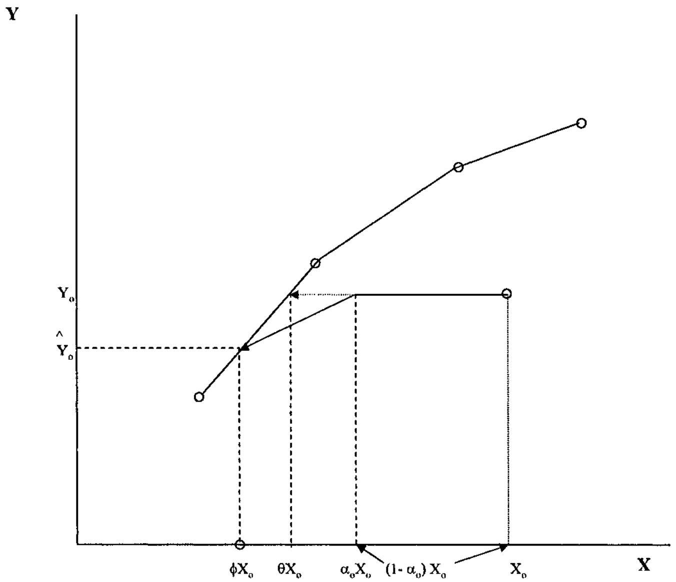

Figure 2-1. Adjusted Projection for an Inefficient DMU

In the situation studied, managers were unable to provide a precise value for γ_j. Rather, they were able to specify this parameter within bounds $\gamma_{1_j} \leq \gamma_j \leq \gamma_{2j}$. Estimation of the ranges $[\gamma_{1j}, \gamma_{2j}]$ posed more difficulty in some patrols than in others. Patrols with relatively uniform traffic and uniform road conditions throughout, presented less of a problem in regard to defining lower and upper bounds on γ. For those patrols where a wide range of circumstances exist among the highway sections making up its network, these ranges were, however, more difficult to capture. In this latter case one finds situations, for example, where a budget reduction can mean that a particular ditching operation to enhance drainage on a small section of the roadway may be shelved. The immediate, or even long term impact of such an activity can be difficult to quantify in terms of road deterioration, etc. Specifically, it can be the case that large budget cuts may effect few drivers, or many, depending upon the type of activity foregone. In such circumstances, management tended to specify a wider range $(\gamma_{1j}, \gamma_{2j})$ than in situations where there was more certainty. Again, we emphasize that the declared ranges are assumed to be good faith declarations, since the *zero-*

sum game environment leaves little room for any given supervisor to exaggerate his/her needs.

For model development purposes in this section we assume that γ_j is a known value. In the following section, we return to the consideration of a range $(\gamma_{1j}, \gamma_{2j})$. Let patrol o be one for which the frontier target of $(1-\theta_o)X_o$ reduction in resources is *not* achievable, but rather there is a declared maximum reduction of $(1-\alpha_o)X_o$, where $\alpha_o > \theta_o$. Formally, the primal linear programming variant (2.4) of the CCR model (2.3) becomes

$$\begin{aligned} &\min\ \phi \\ &\text{subject to:} \\ &\phi X_o - \sum_{j=1}^{n} \lambda_j X_j && \geq 0 \\ &\sum_{j=1}^{n} \lambda_j Y_j && \geq Y_o[1-\gamma_o(\alpha_o - \phi)] \\ &\sum_{j=1}^{n} \lambda_j && = 1 \\ &\lambda_j && \geq 0, \quad j = 1, \ldots, n \end{aligned} \tag{2.4}$$

or

$$\begin{aligned} &\min\ \phi \\ &\text{subject to:} \\ &\phi X_o - \sum_{j=1}^{n} \lambda_j X_j && \geq 0 \\ &-\gamma_o \phi Y_o + \sum_{j=1}^{n} \lambda_j Y_j && \geq Y_o(1-\gamma_o \alpha_o) \\ &\sum_{j=1}^{n} \lambda_j && = 1 \\ &\lambda_j && \geq 0 \quad j = 1, \ldots, n \end{aligned} \tag{2.5}$$

While slacks are not explicitly displayed in (2.5) they do play a role in the application developed herein. More direct reference is made to slacks, and how they are computed later.

Note that the dual form of this is:

$$\begin{aligned} &\max \mu(1-\gamma_0\alpha_o)Y_o+\omega \\ &\text{subject to:} \\ &vX_o-\mu Y_o=1 \\ &\mu Y_j+\omega-vX_j \quad \le 0, \quad j=1,\dots,n \\ &\mu,\gamma \quad \ge 0, \\ &\omega \quad \text{unrestricted} \end{aligned} \tag{2.6}$$

and the resulting equivalent ratio model is:

$$\begin{aligned} &\max \frac{\mu Y_o+\omega-\gamma_o\alpha_o\mu Y_o}{vX_o-\gamma_o\mu Y_o} \\ &\text{subject to:} \\ &(\mu Y_j+\omega)/vX_j \quad \le 1, \quad j=1,\dots,n \\ &\mu,v \quad \ge 0 \\ &\omega \quad \text{unrestricted} \end{aligned} \tag{2.7}$$

It is noted that since output erosion is an inherent feature in *all* DMUs, it would appear that rather than (2.7) the appropriate ratio model should be:

$$\begin{aligned} &\max \frac{\mu Y_o+\omega-\gamma_o\alpha_o\mu Y_o}{vX_o-\gamma_o\mu Y_o} \\ &\text{subject to:} \\ &\frac{\mu Y_j+\omega-\gamma_j\alpha_j\mu Y_j}{vX_j-\gamma_j\mu Y_j} \quad \le 1 \\ &\mu,v \quad \ge 0 \\ &\omega \quad \text{unrestricted} \end{aligned} \tag{2.8}$$

It can be shown, however, that these two formulations are equivalent, as given by the following theorem.

Theorem 2.1: Problems (2.7) and (2.8) are equivalent..

Proof: It is sufficient to prove that at any point $(\hat{\mu},\hat{\omega},\hat{v})$

$$\frac{\hat{\mu}Y_j+\hat{\omega}-\gamma_j\alpha_j\hat{\mu}Y_j}{\hat{v}X_j-\gamma\hat{\mu}Y_j}\le 1$$

if and only if

$$\frac{\hat{\mu}Y_j + \hat{\omega}}{\hat{v}X_j} \le 1$$

Case 1: Assume $(\hat{\mu}Y_j + \hat{\omega})/\hat{v}X_j = 1$. In this case DMU j is a frontier unit, meaning that $\alpha_j = 1$. Hence,

$$\frac{\hat{\mu}Y_j + \hat{\omega} - \gamma_j\alpha_j\hat{\mu}Y_j}{\hat{v}X_j - \gamma_j\hat{\mu}Y_j} = \frac{\hat{\mu}Y_j + \hat{\omega} - \gamma_j\hat{\mu}Y_j}{\hat{v}X_j - \gamma_j\hat{\mu}Y_j} = 1$$

as well.

Alternatively, assume $(\hat{\mu}Y_j + \hat{\omega})/\hat{v}X_j < 1$.

Let θ_j denote the optimal input-oriented DEA score, for example $\theta_j = (\mu * Y_j + \omega^*)/v * X_j > (\hat{\mu}Y_j + \hat{\omega})/\hat{v}X_j$.

It follows that

$(\hat{\mu}Y_j + \hat{\omega})/\hat{v}X_j = \phi_j \le \theta_j = (\mu * Y_j + \omega *)/v * X_j \le \alpha_j$.

Then,

$$\frac{\hat{\mu}Y_j + \hat{\omega} - \alpha_j(\gamma_j\hat{\mu}Y_j)}{\hat{v}X_j - (\gamma_j\hat{\mu}Y_j)} \le \frac{\hat{\mu}Y_j + \hat{\omega} - \phi_j(\gamma_j\hat{\mu}Y_j)}{\hat{v}X_j - (\gamma_j\hat{\mu}Y_j)} = \frac{\hat{\mu}Y_j + \hat{\omega} - \frac{(\hat{\mu}Y_j + \hat{\omega})(\gamma_j\hat{\mu}Y_j)}{\hat{v}X_j}}{\hat{v}X_j - (\gamma_j\hat{\mu}Y_j)}$$

$$= \frac{\hat{v}X_j(\hat{\mu}Y_j + \hat{\omega}) - (\hat{\mu}Y_j + \hat{\omega})(\gamma_j\hat{\mu}Y_j)}{(\hat{v}X_j)^2 - \hat{v}X_j(\gamma_j\hat{\mu}Y_j)}$$

$$= \frac{(\hat{\mu}Y_j + \hat{\omega})[\hat{v}X_j - \gamma_j\hat{\mu}Y_j}{\hat{v}X_j[\hat{v}X_j - \gamma_j\hat{\mu}Y_j]}$$

$$= \frac{\hat{\mu}Y_j + \hat{\omega}}{\hat{v}X_j} < 1$$

Case 2: Assume $\frac{\hat{\mu}Y_j + \hat{\omega} - \gamma_j\alpha_j\hat{\mu}Y_j}{\hat{v}x_j - \gamma_j\hat{\mu}y_j} \le 1$.

Then $\hat{\mu}Y_j + \hat{\omega} - \gamma_j\alpha_j\hat{\mu}Y_j \le \hat{v}X_j - \gamma_j\hat{\mu}Y_j$ or $\hat{\mu}Y_j + \hat{\omega} - \hat{v}X_j \le (\alpha_j - 1)(\gamma_j\hat{\mu}Y_j) \le 0$. So $(\hat{\mu}Y_j + \hat{\omega})/\hat{v}X_j \le 1$.

Hence, the result. QED.

In the section to follow we examine the output deterioration in the context of highway maintenance crew efficiency.

2.4. THE APPLICATION

Referring again to the highway maintenance example, consider the following sample of 14 patrols.

In this example two outputs were chosen to represent the aggregate service performed by maintenance crews.

Table 2-5. Output and Input Data

	Outputs		Inputs	
Patrol#	Size	Traffic Served	Total Expenditure	Average Rating
1	696	39	751	67
2	616	26	611	70
3	456	25	538	70
4	616	31	584	75
5	560	28	665	70
6	446	16	445	75
7	517	26	554	76
8	492	18	457	72
9	558	27	582	74
10	407	18	700	69
11	463	33	630	78
12	350	88	1074	75
13	581	55	1072	74
14	413	24	696	80

Outputs

Size - a measure that is an aggregate or composite of the number of kilometres of paved surface, amount of paved versus gravel shoulders, etc.

Traffic Served - this measure accounts for the average daily traffic and the length of the roadway served.

Two inputs were used in the analysis, namely:

Inputs

Total Expenditure - the annual maintenance budget for the patrol.

Average Pavement Rating - this is a standard indicator per road section (on a 0-100 scale).

Arguably, one might consider treating *average pavement rating* as an ordinal rather than cardinal variable. In this instance, the model of Cook et al. (1993) might aid in deriving projections. It should be pointed out, however, that the rating is established through formal geotechnical data gathering and as such should be treated as quantitative rather than qualitative. With the inherent lack of precision in this measure, a somewhat more formal treatment could involve the imprecise DEA arguments of Cooper, Park and Yu (1999) and Zhu (2003;2004). We have not undertaken this herein. For a more full description of these factors, see Cook et al. (1990).

It is noted that on the input side, the available budget (total expenditure) is clearly a *discretionary* variable, while the road condition, an indicator of the environment in which the patrol operates, is clearly non-discretionary. Arguably, surface maintenance expenditures such as the filling of potholes and sealing of cracks do have a minor impact on the pavement rating (causing it to increase slightly). However, it is not really at the discretion of management to change the pavement condition in any direct way.

As discussed above, the initial analysis of patrol efficiency was conducted here for two primary reasons. First, there was a desire to determine the benchmark crews against which inefficient ones could be evaluated. This provided management with the best and, even more importantly, the worst performers, hence isolating areas where waste existed, and improvements were possible. A second, and related reason for the analysis, was to have a set of measures that could potentially aid in budget setting. Specifically, under various overall provincial highway maintenance budget scenarios, how should allocations to individual patrols be made?

The input-oriented DEA model of Banker et al. (1984) was applied, but restricting the input variable Average Pavement Rating to be nondiscretionary. Specifically, the mixed discretionary/nondiscretionary version of model (2.3) was applied, namely

$$\begin{aligned}
&\min \theta \\
&\text{subject to:} \\
&\theta x_{io} - \sum_{j=1}^{n} \lambda_j x_{ij} - s_i^1 = 0, \qquad i \varepsilon DI \\
&x_{io} - \sum_{j=1}^{n} \lambda_j x_{ij} - s_i^1 = 0, \qquad i \varepsilon NDI \qquad (2.9) \\
&\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^2 = y_{ro}, \qquad r \varepsilon DO \\
&\sum_{j=1}^{n} \lambda_j = 1
\end{aligned}$$

Here, the set of discretionary inputs DI is the budget, and the nondiscretionary inputs, NDI consists of the single variable pavement rating. Outputs are assumed to be discretionary (DO) to the extent that under budget reductions, patrol crews can choose to service the road network in a manner that is below standard. It is noted that we explicitly represent input and output slacks here as s_i^1, s_r^2 respectively. In solving (2.9) we use a 2-stage process wherein the sum of slacks is minimized in stage 2. This unconventional way of handling slacks has some practical merit here in that for example on the input side we are identifying a minimal reduction in resources needed to reach the frontier proper from a frontier extension point.

Table 2-6 presents the projections and efficiency score θ for each of the 14 DMUs. When positive slacks exist they are displayed in brackets. In this example, exactly 7 of the patrols are efficient, both in the radial sense $(\theta = 1)$, and in the CCR-efficient sense, in that all slacks are zero (see Cooper, Seiford and Tone (2000)). The remaining inefficient units are a mix of properly enveloped (DMU#5), and improperly enveloped units (DMUs #7,9,10,11,13,14).

Table 2-6. Efficiency Scores & Projections

DMU	Size	Traffic	Expenditure	Rating	Score
1	696	39	751	67	1
2	616	26	611	70	1
3	456	25	535	70	1
4	616	31	584	75	1
5	560	28	588	75	.883
6	446	16	445	75	1
7	517	26	531	72 (4)*	.958
8	492	18	457	72	1
9	558	27	543	73.75 (.25)	.934
10	536 (129)	29.67 (11.67)	609	69	.870
11	463	33	589	72.67 (5.33)	.935
12	350	88	1074	75	1
13	581	55	855	69.75 (4.25)	.797
14	479.8 (66.8)	24	510	72.26 (7.74)	.733

**Numbers in brackets represent positive slacks. Note, for example, that the road rating for patrol #7 was 76 meaning that a projected value of 72 leaves a slack of 4.*

In attempting to apply the recommended expenditure reductions arising from the efficiency analysis, some (inefficient) patrols found that the projected values could not be achieved. In consultation with head office maintenance management, patrol supervisors provided a minimum budget level that they believed was necessary to maintain the network at a standard, as set by the department. In the case of patrol #5, for example, it was estimated that at most an 8% budget reduction was possible. Beyond this, it was felt that a reduction in maintenance effort would need to occur, and a lower quality of service would be the consequence.

As discussed earlier, an attempt was made to estimate the range $(\gamma_{1j}, \gamma_{2j})$ for the parameter γ_j, for each patrol j. Figure 2-2 illustrates how the erosion projection of Figure 1 might now appear.

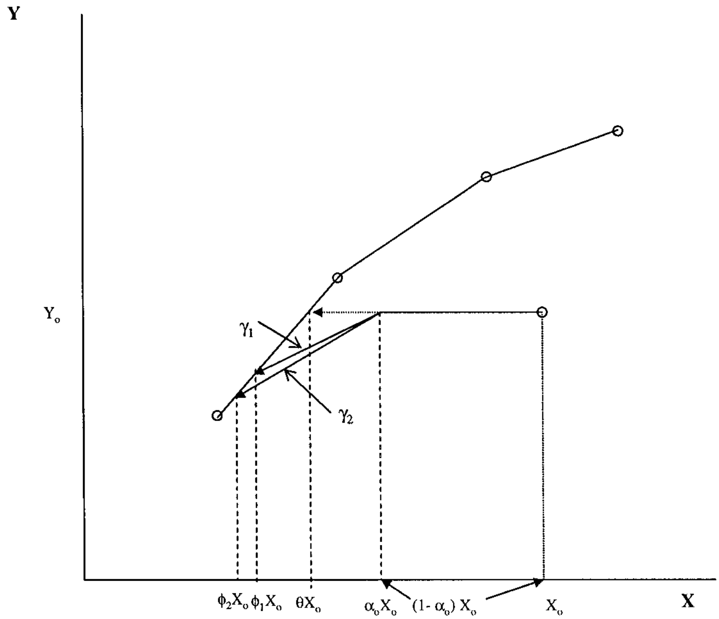

Figure 2-2. Range of Adjustable Projections

Recall that $1-\gamma_j$ is the expected percentage reduction in outputs (service) per radial percentage unit reduction in those discretionary inputs of X_j (i.e., the maintenance budget for j). For discussion purposes here, this range was taken to be $\gamma_j \in [.2,.8]$ for each j. The results for model (2.5) for each of $\gamma_{j1} = .2$ and $\gamma_{j2} = .8$, are displayed in Table 2-7. It is noted that only results for inefficient units are shown since all projections for efficient units are, by definition, the same as their current positions.

For slackless projections such as is the case for DMU #5, projected outputs help to reveal the extent of erosion of the system. Here, under the current status, size and traffic managed are represented by the values (560,28). The computed efficiency score for this patrol is .883, meaning that a reduction in expenditure of 11.7% would be needed in order to reach the frontier of best performance. The projection corresponding to this rating is shown in the row labeled 'Unadjusted.'

In this case, the claimed maximum reduction possible, without eroding outputs, is 8% (α_0 = 92% as compared to θ = 88.3%). Below the 92%

level, if outputs decline at a rate of $\gamma_1 = .2$ (20% of the input reduction beyond that point), then the resulting projected size and traffic that can be serviced are (555.4, 27.8). This represents a .7% decrease in service. Note that the new efficiency score is given by ϕ = .879. The corresponding projection for γ_2 = .8 is (527.7, 26.5), or a 5.7% decrease in outputs, with ϕ=.852. Again, see Figure 2-7.

Table 2-7. Efficiency Scores, Unadjusted and Adjusted Projections

DMU	Status	Size	Traffic	Exp.	Rating	Effic.	α
5	current	560	28	665	70	—	—
	unadj.	560	28	587.4	70	.883	—
	γ_1=.2	555.4	27.8	584.2	70	.879	.92
	γ_2=.8	527.7	26.5	566.8	70	.852	.92
	θ bdd.	543.6	27.2	587.4	70	.883	
7	current	517	26	554	76	—	—
	unadj.	517	26	531	73(3)	.958	—
	γ_1=.2	515.7	25.9	530.2	73(3)	.957	.97
	γ_2=.8	508.9	25.6	526.6	72.9(3.1)	.951	.97
	θ bdd.	512.2	25.8	530.9	73.9(2.1)	.958	
9	current	558	27	582	74	—	—
	unadj.	558	27	543.3	73.8(.2)	.934	—
	γ_1=.2	556	26.9	542.2	73.7(.3)	.932	.95
	γ_2=.8	545.8	26.4	537.0	73.5(5)	.923	.95
	θ bdd.	550.6	26.6	543.3	74.0	.934	
10	current	407	18	700	69	—	—
	unadj.	536(129)	29.7(11.7)	609	69	.87	—
	γ_1=.2	536(136)	29.7(12)609	609	.87	.95	
	γ_2=.8	536(155)	29.7(12.8)	609	69	.87	.95
	θ bdd.	536(155)	29.7(12.8)	609	69	.87	
11	current	463	33	630	78	—	—
	unadj.	463	33	589.3	72.7(5.3)	.935	—
	γ_1=.2	463	33	589.3	72.7(5.3)	.935	.935
	γ_2=.8	463	33	589.3	72.7(5.3)	.935	.935
	θ bdd.	463	33	589.3	72.7(5.3)	.935	
13	current	581	55	1072	74	—	—
	unadj.	581	55	854.8	69.7(4.3)	.797	—
	γ_1=.2	573.1	54.3	839	70.5(3.5)	.783	.85
	γ_3=.8	500	47.3	724.2	73.2(.8)	.676	.85
	θ bdd.	556.6	52.7	854.8	73.2(.8)	.797	
14	current	413	24	696	80	—	—
	unadj.	479.8(66.8)	24	509.9	72.3(7.7)	.733	—
	γ_1=.2	481(78.3)	23	504.6	72.2(7.	.725	.85
	γ_3=.8	486(124.1)	21	483.6	72.1(7.9)	.695	.85
	θ bdd.	436(62)	21.8	509.9	74.6(5.4)	.733	

Thus, under the worst case scenario, patrol 5 could experience a 5.7% decrease in service delivered to the road user and to the tax-paying public. Recall that while decreased service can take several forms, it is useful to view this scenario as portraying a lower quality product, a faster deterioration of the network, and a higher capital expenditure in the long run.

For projections with slack on the output side, a slightly different interpretation takes place. Consider the two situations portrayed by patrols #10 and #14. For #10, the projected outputs are the same under all three scenarios (unadjusted, γ_1 and γ_2). For example, the frontier projected size is 536 in all situations, and the efficiency score remains at 87%. The actual projected point (on the frontier extension)is, however, given by

Frontier projection-slack
= 536-129 = 407 in unadjusted case
= 536 -136 = 400 in γ_1 case
= 536 - 155 = 381 in γ_2 case.

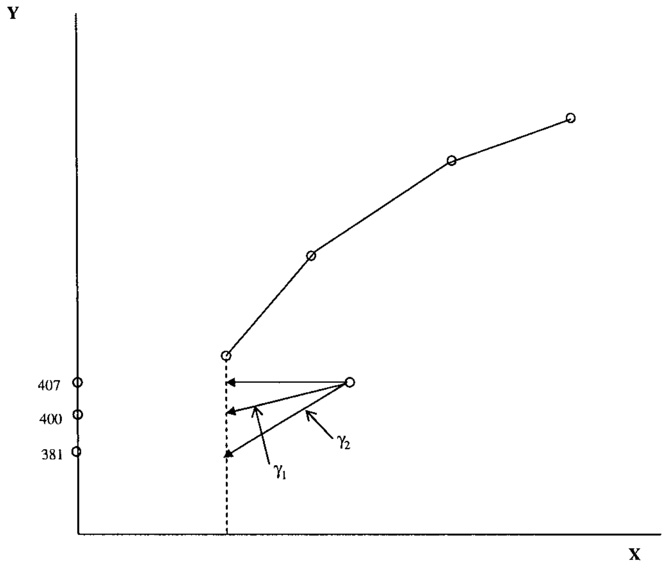

Figure 2-3. Projection with Output Slack

Figure 2-3 provides a representation of this phenomenon. (Note that a similar result occurs for the traffic factor). From a lost service perspective, it is these frontier extension values that are of interest to management.

For #14, the situation is very similar except that there is slack in only one of the outputs (size), and the efficiency score continues to decrease as we move from the unadjusted projection where θ =.733, to γ_1 (ϕ = .725) and to a_2 (ϕ = .695).

2.4.1 Base-Line Budget Considerations

The rationale for deriving input-oriented efficiency measures in the present setting, appears to be twofold. First, the measures point to those patrols that are inefficient, and those that are efficient; this sets out benchmarks that management can utilize to help poorly performing patrols to improve their status. Second, efficiency measures can aid in setting budgets. Budget planning here would appear to be an exercise in *scenario* analysis, and the results obtained from Tables 2-6 and 2-7 put bounds on the *minimal* fiscal requirements for the maintenance function. One scenario is that provided by the achievable projections described by the α_j measures. Specifically, the $(1-\alpha_j)$ % reduction in discretionary inputs (maintenance expenditure, in this case), can be achieved without any erosion to output measures. Under this scenario, for the sample of 14 patrols considered, the current budget of \$9359 could be reduced to \$8874. Thus, a budget reduction of \$485 (thousand) would appear to be immediately achievable.

The minimal budget projections under the γ_1 and γ_2 output erosion scenarios are given by \$8656, and \$8494 respectively. These lower anticipated budgets, depending on the outcome erosion rates that may result, provide management with a guide as to the possible savings obtainable if all DMUs were required to move to a frontier efficiency status.

Possibly, a more realistic and fair system of minimal budget setting would be one wherein patrols are required to reduce expenditures only by the original $1-\theta$ measure. Specifically, if no output erosion occurred, an inefficient patrol 0 would need to operate only at an expenditure level of θx_{10}, to be deemed efficient, rather than at the often lower level of ϕx_{10}. Here, x_{io} denotes the expenditure level (i=1) for DMU o. For example, in the case of patrol 13, the budget allocation would be .797 × 1072 = \$854.8 (thousand), rather than the lower figures \$839 and \$724.2 corresponding to γ_1 and γ_2, respectively. To compute the output erosion corresponding to this more favorable $0ax_{10}$ position, we resolve a modified version of (2.5) wherein ϕ is restricted to not be less than θ. Figure 2-4 illustrates this idea.

The resulting projections are shown in Table 2-7, corresponding to the status entitled θ-bdd. In computing these projections the most pessimistic view of output deterioration has been assumed (γ_2=.8 was used). Except in cases #10 and #11, the projections for inefficient units are not on the frontier, but such units would be operating at budget levels that would normally be seen as more appropriate than those resulting from the γ_1, γ_2 scenarios. The overall minimal budget for the 14 patrols in this case is \$8682. Let us regard this as a *base-line* or starting budget position.

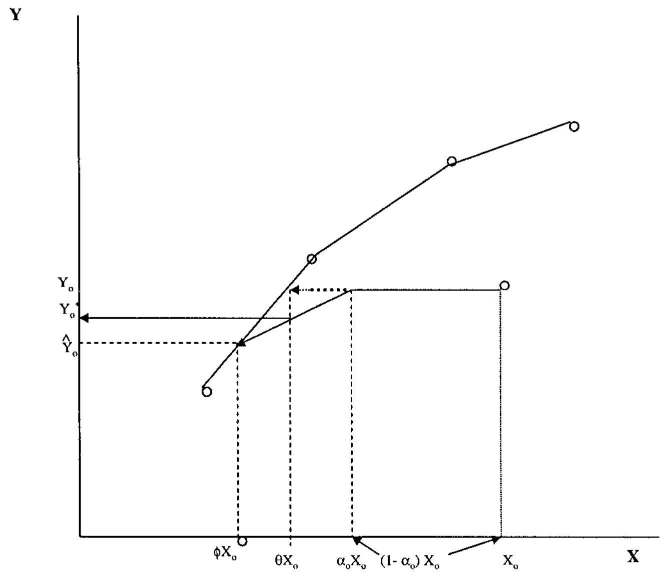

Figure 2-4. θ-Projection for an Inefficient DMU

2.4.2 Budget Allocation Beyond the Base Line

The various projections discussed above provide management with a broad scope for making budget decisions. Let us assume that the supplied γ_j (or expected γ_j) represent reality and are not exaggerated claims by the management of DMU $_j$. If the organization adopts the θx_{io} position as a form of base-budget status, then the aggregate base budget operating level is

$$B_b = \sum_{j=1}^{n} \theta_j x_{ij} \tag{2.10}$$

At this base budget level, patrol j would be providing a level of service of

$$y_j^* = y_j[1-\gamma_j(\alpha_j - \theta_j)], \tag{2.11}$$

if j is experiencing output erosion (i.e., $\theta_j < \alpha_j$). Otherwise, $y_j^* = y_j$.

One advantage of adopting a base-budget approach as the starting point for allocating maintenance funding to patrols, is that it becomes somewhat transparent as to what budget impacts will be for funding *above* the base level. For example, if there is a \$1 (thousand) increase in patrol js budget above the θx_{ij} level, one can estimate the increase in y_j that can be expected to occur. Specifically, the improved values of the components of y_j, currently at y_j^*, are given by:

$$\begin{aligned} y_j' &= y_j[1-\gamma_j(\alpha_j-\theta_j\frac{1}{x_{1j}})] \\ &= y_j^* + \frac{\gamma_j y_j}{x_{1j}} \end{aligned} \tag{2.12}$$

Note that y_j/x_{1j} is the vector of existing output rates (outputs per monetary unit of budget), and γ_j is the expected rate of increase in outputs per monetary increment to the base budget.

In the single output case (y_j is a scalar), one could allocate additional resources to patrols according to the per unit *gain factor* $\gamma_j y_j/x_{1j}$ (ranked in descending order). Specifically, if DMU j_1 has the highest gain factor, then one would presumably increase patrol j_1 s budget by δ_{j1} so that

$$\delta_{j1}\gamma_{j1}y_{j1}/x_{1j1} = y_{j1} - y_{j1}^*$$

or

$$\delta_{j1} = [(y_{j1} - y_{j1}^*)x_{1j1}]/(\gamma_{j1}y_{j1}) \tag{2.13}$$

If resources still remain, allocate funds accordingly to the patrol j_2, whose gain factor is ranked in second place, and so on.

In the multiple output case, optimization is problematic in that the patrol most desirable for a funding increment in regard to the system size dimension, may not rank highest on the traffic dimension. Thus, the problem is multi-criteria in nature, with a ranking of the patrols being available for each output type. Since the units that define the outputs are not comparable, one reasonable mechanism for ranking the patrols (for consideration for budget increments) would be to replace the vector y_j by the weighted aggregate output $\mu_j y_j$, where μ_j is the optimal multiplier vector (shadow prices from (2.9) for problem j).

Pure optimization here may be somewhat elusive in that γ_j, as discussed earlier, is known only within a range $(\gamma_{1j}, \gamma_{2j})$. Management would need to choose an appropriate value γ_j in this range if a comparison of patrols is to be made.

2.5. DISCUSSION

This chapter has examined the application of DEA in the area of highway maintenance. It has illustrated as well, the difficulty of matching theoretical and achievable targets.

The suggested modifications to the conventional DEA model help to capture the *consequences* on the output side that can occur when inputs are reduced according to the computed performance measures. The failure to realize projected reductions in resources without such consequences in many real world settings can, in most instances, be attributed to factors not included in the modeling exercise. These factors commonly pertain to the *environment* that one DMU may face versus that of its peers. This environment may be physical (differences in road sub-surface structures in maintenance patrols, for example,) or demographic (e.g., customer mix characteristics in financial services settings). Another explanation relates to the *random* nature of outputs or input requirements. In the maintenance crew setting, annual maintenance needs on highways (i.e., budget requirements) are greatly a function of weather, severity of winters, and so on. Geographical location plays an important part. It can be that frontier DMUs are those located in geographically favorable settings, where winter maintenance needs are minimal and roadway deterioration is less prevalent than in other areas. Thus, maintenance needs are random and frontier DMUs can be outliers at the lower tail of the maintenance cost distribution.

Earlier attempts to introduce categorical variables to permit comparison of a DMU to only those others that are proper peers, did not seem to resolve or explain the gap between theoretical and achievable targets. This necessitated the application of model (2.5). This model will hopefully provide a useful enhancement to the existing DEA methodology. It provides a bridge between theoretical performance targets and the practical situations facing DMU management.

REFERENCES

1. Banker, R.D., A. Charnes and W.W. Cooper (1984), Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science* 30, 1078-1092.
2. Banker, R.D. and R.C. Morey (1986), The use of categorical variables in data envelopment analysis, *Management Science* 32, 1613-1627.
3. Charnes, A., W.W. Cooper and E.L. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2, 429-444.
4. Cook, W.D., Y. Roll, and A. Kazakov (1990), A DEA model for measuring the relative efficiency of highway maintenance patrols, *INFOR* 28, 113-124.
5. Cook, W.D., Y. Roll, L. Seiford and A. Kazakov (1991), A data envelopment approach to measuring efficiency: case analysis of highway maintenance patrols, *Journal of Socio-Economics* 20, 83-103.

6. Cook, W.D. and J. Zhu (2001), Output deterioration with input reduction in data envelopment analysis, *IIE Transactions*, 35, 309-320.
7. Cook, W.D., A. Kazakov, and Y. Roll (1994), On the measurement and monitoring of relative efficiency of highway maintenance patrols, Data Envelopment Analysis: Theory, Methodology and Applications, Charnes, A., Cooper, W.W., Lewin, A. and Seiford, L. (eds.), Kluwer, New York. NY.
8. Cook, W.D., Kress, M., and Seiford, L.M. (1993), One the Use of ordinal Data in Data Envelopment Analysis, *Journal of Operations Research Society* 44, 133–140.
9. Cooper, W.W., Park, K.S., and Yu, G., (1999), IDEA and AR-IDEA: Models for Dealing with Imprecise Data in DEA, *Management Science* 45, 597–607.
10. Cooper, W.W., Seiford, L., and Tone, K., (2000), Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-solver Software, Kluwer Academic Publishers, Boston.
11. Deller, S.C. and C.H. Nelson (1991), Measuring the economic efficiency of producing rural road services, *American Journal of Agricultural Economics* 73, 194-201.
12. Roll, Y., W.D. Cook and B. Golany (1991), Controlling factor weights in data envelopment analysis, *IIE Transactions* 23, 2-9.
13. Rouse, P., M. Putterill and D. Ryan, (1997), Towards a general managerial framework for performance measurement: A comprehensive highway maintenance application, *Journal of Productivity Analysis* 8, 127-149.
14. Rousseau, J., and J. Semple (1993), Categorical outputs in data envelopment analysis, *Management Science* 39, 384–386.
15. Shereman, H.D., and G. Ladino, (1995), Managing bank productivity using data envelopment analysis (DEA), *Interfaces* 25, 60–73.
16. Zhu, J. (2003), Imprecise data envelopment analysis (IDEA): A review and improvement with an application, *European Journal of Operational Research* 144, 513-529
17. Zhu, J. (2004), Imprecise DEA via standard linear DEA models with a revisit to a Korean mobile telecommunication company, *Operations Research* 52, 323-329

This chapter is based upon W.D. Cook, Y. Roll and A. Kazakov, 1990, A DEA Model for Measuring the Relative Efficiency of Highway Maintenance Patrols,INFOR, 28, No. 2, 113-124, and *W.D.* Cook. and J. Zhu, 2001, Output Deterioration with Input Reduction in Data Envelopment Analysis, *IIE Transactions*, 35, 309-320, with permission from INFOR and Taylor & Francis, respectively.

Chapter 3

PRIORITIZING HIGHWAY ACCIDENT SITES

3.1. INTRODUCTION

The problem of designing a safety retrofit program involves several components: (1) forecasting target accidents at a selected set of road sites, (2) evaluating the effectiveness of various measures for reducing accidents, and (3) prioritizing the sites in order of effectiveness. A significant body of literature has been dedicated to the first two components, and in particular, to accident prediction modeling. In the present chapter, however, we do not address these two components, but rather we concentrate on the third component, namely, the prioritizing of sites. The discussion herein is based upon Cook, Kazakov and Persaud (2001).

In practice, prioritization of accident sites has been dealt with in various ways:

- rank ordering by total target accident counts;
- converting all target accidents to the equivalent of accidents that involve only property damage and then ranking sections according to these equivalent accidents;
- ranking according to the ratio of the benefit (reduction in accidents) to the cost of applying the recommended retrofit measures.

While these methods for ranking accident sites all have merit, they generally fail to recognize the multi-criteria nature of the problem at hand.

In this chapter we present a DEA-based procedure for selecting a retrofit program in a budget constrained environment. The procedure specifically acknowledges the fact that multiple criteria or factors are involved, for example, different accident classes, and different kinds of costs, such as government agency- and user costs. Furthermore, the method adopted takes account of the fact that there is no well established means of converting these multiple criteria to a single dimensional problem.

3.2. THE PROBLEM

Suppose that a set of I road sections of a given length, for example, 0.5 km have been identified for possible safety treatment. Consider for each road section $i \in \{1,\ldots,I\}$ a finite set J_i of potential retrofit measures. Members of J_i may be combinations of individual retrofit measures, for example, illumination together with shoulder widening. It is understood that for each measure j, the *target accidents* can be identified and measured on each section. For example, for roadway illumination, target accidents would be some portion of all those accidents occurring at night. With this definition, a particular accident might be the target for more than one retrofit measure, and it is possible that some accidents may not be targeted at all.

Let t_{ijk} be the *expected* number of target accidents of severity k for retrofit measure j on road section i. A point estimate $\hat{t}_{ijk}$ of t_{ijk} can be obtained from any one of a number of forecasting techniques. Here, we use the Empirical Bayesian procedure, (see Hauer (1986) and Persaud (1995)), which makes use of the *actual* number of such accidents for this site in the recent past. We do not concentrate here on the accident prediction aspect of designing a retrofit program, but rather refer the reader to the relevant literature on forecasting techniques. See, for example, Abbess et al. (1981), Hauer (1992) and Persaud (1993). It is pointed out that the statistical distribution of t_{ijk} can also be estimated to reflect the uncertainty in the point estimate $\hat{t}_{ijk}$.

For each retrofit measure, and each accident severity class, there is an accident modification or *reduction factor* a_{jk} such that the expected annual number of target accidents in class k after the implementation of the retrofit j on section i is $a_{jk}t_{ijk}$. Thus, the point estimate of the annual *accident savings* is $y_{ijk} = t_{ijk}(1-a_{jk})$.

The application of a retrofit measure to a road section, therefore, gives rise to a *set* of *outputs* or benefits $\{y_{ijk}\}_{k=1}^{K}$. Thus, retrofit measures have multidimensional rather than single dimensional outcomes or consequences. The current state of practice is to attach weights u_k to the different accident types k, thereby converting the multiple dimensional outcomes y_{ijk} to a uni-dimensional benefit $\sum_{k=1}^{k} u_k y_{ijk}$. In principal, the u_k should reflect the saving achieved *per prevented accident* of type k. As an example, in the state of Kentucky weights have been selected to be representative of insurance cost relating to such accidents. In that jurisdiction, all accidents are grouped into three major categories—(1) fatalities, (2) injuries, and (3) property damage only. Empirical studies carried out there concluded that the relative weights 9, 5 and 1 on categories (1), (2) and (3) respectively, reasonably characterize the respective costs. See Troxel (1993). Studies carried out in other jurisdictions have lead to somewhat different weights. Thus, the relative values or prices $\{u_k\}_{k=1}^{K}$ associated with these outcomes (for example, the accident costs), are not well established in the strict sense, although estimates of average accidents costs are available in various jurisdictions. To complicate matters further, there are also multidimensional inputs associated with the application of any retrofit treatment; two of the obvious ones are (1) the cost x_{ij1} incurred by the transport ministry (hence, the tax paying public) of actually applying the measure, and (2)costs x_{ij2} experienced by drivers that are due to lost time with road closures. The transport ministry cost associated with any retrofit measure can be directly stipulated; user or driver costs are less specific. A possible surrogate for driver cost or inconvenience is $x_{ij2} = ADT \times t$, where t is the number of days during which the repair is underway and ADT is the average daily traffic. This definition of driver cost or inconvenience is simply an estimate of the number of drivers being exposed to the repair operation. Arguably, a more suitable reflection of driver inconvenience should take account of peak period traffic and average trip delay per driver. Until suitable data is available, however, the above definition of x_{ij2} will be applied, and is the common one adopted in practice.

Obviously, other inputs might take the form of *environmental* variables such as adjustments to ministry costs. Such adjustments could, for example, capture characteristics of the jurisdiction where the repair is being carried out; an example might be the cost of transporting materials to the repair site in one district versus another. For purposes herein, however, we use as inputs only the two costs indicated above.

Stated in simple terms, the problem of selecting a best set of road sections for safety improvement is one of finding those sections that yield the greatest accident reduction benefits $\{B_{ij}\}$ at the lowest (combined agency- and user) costs $\{C_{ij}\}$. In purely technical terms, B_{ij} and C_{ij} can be expressed in the additive forms

$$B_{ij} = \sum_{k=1}^{K} u_k y_{ijk}$$

and

$$C_{ij} = \sum_{r=1}^{R} v_r x_{ijr},$$

for some sets of multipliers $\{u_k\}_{k=1}^{K}$ and $\{v_r\}_{r=1}^{R}$. As indicated above, the u_k represent the relative weights or importance attached to the various accident classes; the most tangible definition of importance is the public cost associated with those classes. Similarly, the v_r are the relative importance multipliers for the inputs, which in our example are transport ministry costs and user costs. How these values should be chosen is less transparent than is true of the u_k.

In some transportation departments no attempt is made to consider any input (cost), aside from the direct expense to the department; specifically, only x_{ij1} is considered. In other situations such as is the case for the Ontario ministry, studies have attempted to capture the cost per driver, per hour lost in travel time, as a measure to total driver inconvenience. This measure then becomes the exchange rate between one hour lost by the driver versus \$1 expended by the transportation department. This then dictates v_1 and v_2. As with the output side, however, there can be a significant degree of variability in this exchange rate, depending, for example, on the composition of trucks versus private automobiles on the road section i in question. As well, the amount of business travel as compared to other travel influences this rate. So, arguably v_1 and v_2 may be estimated in some range, but would be difficult to fix precisely.

Clearly, if one could derive aggregate values B_{ij} and C_{ij}, the selection problem then could be viewed in terms of finding those sections i whose (benefit/cost) ratios B_{ij}/C_{ij} are largest. In purely economic terms we would be choosing accident sites where the payoffs in accident reductions are greatest relative to the monetary (agency- and user) investments. There are several problems, however, with this approach, with the principal one being that there is no correct set of multipliers $\{u_k, v_r\}$. Furthermore, one needs to evaluate the relative worths of various retrofit measures for each site. In the section to follow we present a model for dealing with these problems.

3.3. APPLICABILITY OF THE DEA METHODOLOGY

In some respects, the problem of selecting accident sites can be viewed in the context of multi-attribute or multi-criteria decision making (MCDM). A vast literature exists on MCDM, which is covered extensively in Cook and Kress (1992). One particular area of MCDM is multi-attribute utility theory. Utility models attempt to derive a function which transforms a set of non-comparable attributes for an entity (e.g., an accident site), into a single value; a multi-dimensional problem is thus converted to a single dimensional problem. One form of utility function views attributes as additive and linear. The function, therefore, is comprised of a set of weights which when applied, say, to different accident types would derive a uni-dimensional value measuring the overall benefit from safety improvements applied to any given accident site.

While conceptually utility theory is a viable means of deriving an overall measure for site improvements, it does have certain disadvantages. First, the model provides for a single set of weights which would apply to all accident sites. There are generally no clear rules for how such weights should be chosen, and, as well, one may arguably want different weights for different sites, which could make allowance for differences pertaining to types of drivers, roadside conditions and so on. Second, some attributes can be seen as outputs from safety improvements (accident reductions) and others as inputs to the improvement process (e.g., safety expenditures). Utility theory has no convenient way of allowing for this dichotomy.

DEA is a tool which is, in certain respects, an extension of the utility theory model. It views data factors as being separated into two groups, and as well, allows for different multipliers for different accident sites. How does DEA relate to the discussion regarding B_{ij} and C_{ij} above?

Referring again to the discussion in the previous section, it can be argued that while viewing the relative desirability of accident sites in terms of a benefit/cost ratio is an appropriate way to proceed, the concept presumes the existence of fixed weights. So, while the expression $9y_{ij1} + 4y_{ij2} + 1y_{ij3}$, for example, captures the aggregate benefit (in the Kentucky model) if the weights were *accurate*, it is more correct to write the expression as $(9 \pm \Delta_F)y_{ij1} + (5 \pm \Delta_I)y_{ij2} + (1 \pm \Delta_p)y_{ij3.}$ Here, $9 \pm \Delta_F$, for instance, reflects the fact that the importance attached to a fatal accident (insurance cost) lies in the range $(9 - \Delta_F)f, 9 + \Delta_F)$, where Δ_F could be obtained from data on past fatality settlements.

What is required, therefore, is a mechanism for applying benefit/cost analysis in this broader context, taking into consideration the uncertainty relating to the multipliers. The DEA methodology of Charnes et al. (1978) was designed specifically as a tool for evaluating different entities (for example, accident sites), wherein there is inherent uncertainty as to the values of the multipliers of inputs and outputs. Specifically, it provides a basis for assigning appropriate multipliers to facilitate benefit/cost analysis in this more general setting.

One possible criticism of the DEA approach might be that too much flexibility is permitted in the choice of the weights. In particular, an accident site that has no fatalities or injury accidents may still receive a high rating z simply because a low or even zero weight can be placed on those two accident classes and a large weight may be placed on property damage accident reductions. Clearly, such a choice for weights is not consistent with what is known to be appropriate for describing the relative importance of the various classes of accidents. Specifically, this may result in the selection of sites for improvements where only property damage accidents have happened instead of truly hazardous locations where fatalities have occurred, and meaningless prioritization could result. To prevent undesirable choices of multipliers, a modified version of model (1.3) of Chapter 1 can be employed wherein restrictions on the weights can be imposed. Many different versions of weight restrictions have been examined in the literature, including *absolute* lower and upper limits on individual multipliers, as discussed in Cook et al. (1990), and in Chapter 2, the *assurance region* method of Thompson et al. (1992), and the *cone-ratio* method of Charnes et al. (1990). The latter may be appropriate here, in that it permits one to impose upper and lower bounds on ratios of weights. For example, if it is felt that the public cost of a fatality at any given site will be at least 4 times that of an injury accident, but not more than 9 times, then if u_1 and u_2 are the multipliers for fatalities and injuries, respectively, a restriction of the form

$$4 \le \frac{u_1}{u_2} \le 9$$

can be imposed. This reduces of course to the linear constraints

$$u_1 - 4u_2 \ge 0 \quad \text{and} \quad u_1 - 9u_2 \le 0.$$

We point out that the range (4,9) is purely for illustrative purposes herein. The choice of range would, in practice, be jurisdiction specific. Similar restrictions may be selected to control the relative sizes of the multipliers v_1 and v_2 on ministry and driver costs.

In the section to follow we apply the DEA methodology to a sample of road sections in Ontario where accidents have occurred in the past.

3.4. APPLICATION TO A SAMPLE OF SAFETY SECTIONS

For purposes of demonstrating the DEA tool in this setting, a sample of 42 road sections was selected for analysis. Table 3.1 displays the data. The data in the first three columns represent the estimated reductions in numbers of accidents that will occur if the retrofit treatment is undertaken. These numbers are based upon accident reduction factors available in the literature applied to projected numbers of target accidents likely to occur on the sections. It is to be noted that the actual data has been scaled for purposes of analysis. Specifically, the actual figures for the first road section were in fact

Fatality	Injury	PDO	Cost	Traffic
0.0034	0.041	0.520	100	4090

These were scaled to 34,41,520,100 and 409 respectively. Since the DEA model is scale invariant, this transformation does not affect the analysis. It is noted that rows 1–5 in Table 3-1 are in fact the same section with five different retrofit measures and associated accident reduction and agency cost factors. Similarly, rows 6–10 represent a particular road with five different treatments. For each of the other road sections, only a single retrofit measure is considered.

Two separate analyses were carried out utilizing the DEA package discussed in Chapter 1 and contained herein. In the first analysis, no cone-ratio bounds were placed on the accident type multipliers (i.e., the multipliers u_1, u_2 and u_3 were left unrestricted in terms of upper and lower bounds). Table 3-1 displays the DEA scores, labeled Theta I. In the second analysis limits were imposed of the form

$$4 \leq \frac{u_{fatality}}{u_{injury}} \leq 10$$

$$3 \leq \frac{u_{injury}}{u_{PDO}} \leq 8,$$

and the resulting DEA scores are shown as Theta II.[1]

3.4.1 Selecting Treatments and Sections

The problem to be addressed is one of choosing those safety initiatives that should be undertaken within budget restrictions, and at the same time selecting an appropriate treatment for each chosen section. One approach to this problem is to rank the (section, treatment)- combinations in order of their aggregate benefit/cost ratios. The Theta-parameters provide these ratios.

A first step is to rank the impact of the various treatments for any given section. This means, for example, that we would take the five treatments for the first section, and select the highest ratio. (For example, choose the highest among the first five numbers under Theta I in Table 3-1. Clearly, the second, fourth or fifth would all qualify as recommended treatments.)

The second step is to rank order the resulting Theta-scores (having chosen one for each section), then choose the first K sections such that the corresponding K costs fall within the budget, but if a $K+1$ st were included, the available budget would be exceeded.

[1]It is noted that because of the scaling of the actual data presented earlier, the bounds applied to the scaled data were in fact. $0.4 \leq u_1/u_2 \leq 1$ and $3 \leq u_2/u_3 \leq 8$.

Table 3-1. Fatality data for various road sections

Section	Fatality	Injury	PDO	Cost	Traffic	Theta I	Theta II
1	34	41	521	100	409	0.83593	0.77533
2	13	79	8	50	409	1	0.46273
3	40	57	362	60	409	0.93450	0.81590
4	153	216	1108	450	409	1	1
5	140	137	1100	400	409	1	0.94432
6	42	22	536	100	872	0.50499	0.46394
7	56	4	144	60	872	0.60523	0.25174
8	43	59	478	200	872	0.40190	0.38611
9	86	73	895	110	872	0.86070	0.84261
10	29	69	751	50	872	1	0.91083
11	9	1	204	60	614	0.25573	0.22808
12	40	16	748	200	910	0.53988	0.44476
13	47	2	309	50	587	0.70954	0.47937
14	58	40	59	100	337	0.92708	0.42128
15	3	25	572	100	860	0.48858	0.44707
16	79	65	749	50	866	1	1
17	11	28	310	200	861	0.24489	0.21999
18	4	14	900	40	685	1	1
19	101	95	819	110	648	1	1
20	28	86	223	160	391	0.75470	0.51924
21	87	93	1050	300	1047	0.64038	0.60160
22	61	21	219	40	531	1	0.58744
23	46	35	17	60	592	0.64865	0.28478
24	23	31	189	50	528	0.49614	0.41923
25	9	24	715	50	918	0.71027	0.69033
26	52	17	274	100	436	0.70221	0.45303
27	42	26	352	40	512	0.78433	0.71331
28	14	8	687	200	725	0.57809	0.43687
29	4	26	462	40	511	0.79083	0.72433
30	10	37	406	100	978	0.38437	0.34152
31	14	65	139	100	366	0.73879	0.45234
32	37	55	143	100	708	0.49837	0.32042
33	55	70	158	100	748	0.65832	0.39811
34	18	57	657	100	780	0.68541	0.62642
35	35	26	636	100	865	0.56140	0.53051
36	24	45	507	100	954	0.48131	0.44618
37	36	65	412	100	946	0.54942	0.45745
38	5	16	138	100	478	0.21476	0.19328
39	50	61	161	100	759	0.58011	0.36375
40	39	59	782	100	769	0.78650	0.75580
41	53	43	99	100	811	0.49538	0.27450
42	27	11	162	100	522	0.32131	0.23140
43	14	72	76	100	595	0.59069	0.30209
44	20	4	183	100	369	0.33563	0.27251
45	49	28	482	100	1123	0.42472	0.39457
46	48	27	241	100	658	0.49240	0.34601
47	2	26	743	100	962	0.56746	0.52569
48	29	31	119	100	514	0.37126	0.26700
49	21	22	223	100	700	0.25822	0.25463
50	18	25	488	100	502	0.64175	0.57059

Recall that the imposition of cone-ratio bounds has the effect of reducing the Theta parameters. Consequently, fewer "ones" appear under the Theta II column in Table 3-1. Thus, appropriate bounds allow one to break ties. Consider, for example, the second treatment for the first safety section. Under Theta I this section was rated 1 primarily because a heavy weight was placed on the injury accident figure of 79, and very little weight was given to the fatality figure 13. For treatment 14, the opposite is true (fatalities are weighted high and injuries low.) With the imposition of the cone-ratio bounds, the weight on fatalities is forced to be at least 4 times greater than that on injuries, and so on. As a result, in the second run (Theta II), the rating on the second treatment dropped to .46. Thus, treatment 14 would now be chosen for the first section, primarily for its dominance in reducing fatalities.

Applying this logic using Theta II, and with a budget of \$1100K or \$1,100,000, the chosen sections and treatments would be:

Section	Theta II	Cost (K$)
4	1.00	450
16	1.00	50
18	1.00	40
19	1.00	110
10	0.91	50
40	0.75	100
29	0.72	40
27	0.71	40
25	0.69	50
34	0.62	100
21	0.60	55

In this case, the list of projects consumes \$1085K of the \$1100K, and no other project exists which can be accomplished using the remaining \$15K. In practical terms, one might argue that some other selection of projects might have been chosen which could consume a larger portion (perhaps all) of the given budget of \$1100K. If, for example, only a budget of \$1000K had been available, only those sections down to DMU 25 would have met the strict criterion of falling within the budget limit. That is, sections $4,16,18,\ldots,25$ consume \$930K, and if we go to DMU 34 with its cost of \$100K, we would run over the budget limit. Of course, if one ignores DMU 34 altogether and goes to the next section on the list, i.e., DMU 21 with a cost of \$55K, then it could be included in the set. Thus, refinements to the basic ranking idea are easily implemented.

Obviously, a less ad hoc approach to the ranking method is to attempt to apply a standard knapsack algorithm. This is in recognition of the fact that the problem we face here is really a capital budgeting problem with the objective being to choose those projects with the highest ratings (Thetas), and with a single constraint on the cost. One is reluctant to take this approach here, however, since in the usual capital budgeting problem, the objective function is generally one where meaningful numerical data such as profit or revenue is being maximized. The Theta values in the present problem are not profits or revenues, but rather are ratings that have come about by way of a process that is completely disconnected from the budget constrained setting that we eventually come to.

An alternative, but somewhat more complex approach, to the resource allocation problem involved with selecting safety projects is that suggested by Cook and Green (2000). Their approach effectively amounts to (implicitly) looking at all *subsets* of projects, each of whose total budget comes as close as possible to the given budget without exceeding it. They use a mixed integer programming technique to search through the various candidate subsets. The benefit of this more involved approach is that it determines that set of projects whose *aggregate* benefit (total reduction in accidents) per dollar spent is maximized. We have not applied this approach to this particular data set.

3.5. CONCLUSIONS

In this chapter a procedure has been presented for selecting a set of safety retrofit projects. One of the complexities surrounding this selection is the multi-dimensional nature of the problem. Specifically, one must consider various accident types on the benefit or output side, as well as agency cost, user inconveniences and possibly environmental factors on the input side. The data envelopment analysis method is applied to this multiple criteria setting using a sample of road sections, each with accompanying proposed retrofit measures.

It should be emphasized that the DEA model structure has been extended in the literature, and would permit a much broader analysis of accident sites than might appear to be the case from the above example. Clearly, a complete analysis of sites should attempt to address the many behavioral factors that can influence accident occurrences—age of drivers, gender split, extent of alcohol involvement, speeds involved, and so on. While no provision to examine these issues was made in the example herein, one could incorporate *qualitative* data factors (see Cook et al. (1992)) as well as *non-discretionary* variables as per Banker and Morey (1986), to handle such non-controllable inputs as average driver-age and speed. Driver mix would be viewed as non-discretionary in that it, unlike economic factors, cannot generally be changed. Such behavioral or driver-mix data may not be available in many jurisdictions, although all accident related information (alcohol involvement, driver age, etc.) are normally contained in police records.

Only three accident categories and two types of costs are used for demonstration purposes herein. Further work is required to enlarge this variable set to include other factors which may be pertinent to the analysis. In particular, factors that more accurately capture roadway user cost, rather than using traffic only, should be considered.

REFERENCES

1. Abess, C., D. Jarrett, and C. Wright, (1981), Accidents at blackspots: estimating the effectiveness of remedial treatment with special reference to the 'regression-to-the-mean' effect, *Traffic Engng Control* 22, 535-542.
2. Banker, R.D. and R.C. Morey (1986), The use of categorical variables in data envelopment analysis, *Management Science* 32, 1613-1627.
3. Charnes, A., W.W. Cooper and E.L. Rhodes (1978), Measuring the efficiency of decision making units. *European Journal of Operational Research* 2, 429-444.
4. Charnes, A., W.W. Cooper, Z.M. Huang and D.B. Sun (1990), Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks, *Journal of Economics* 46, 73-91.
5. Cook, W.D., A. Kazakov and B.N. Persaud (2001), Prioritizing highway accident sites: A data envelopment analysis, *Journal of the Operational Research Society* 52, 303-309.
6. Cook, W.D., and M. Kress (1992), Ordinal Information and Preference Structures: Decision Models and Applications. Prentice-Hall, Englewood Cliffs, NJ.
7. Cook, W.D., Y. Roll and A. Kazakov (1990), A DEA model for measuring the relative efficiency of highway maintenance patrols, *INFOR* 28, 113-124.
8. Cook, W.D. and R.J. Green (2000), Project selection using DEA, *Socio Economic Plan* 34, 85-99.
9. Hauer, E. (1986), On the estimation of the expected number of accidents. *Accident Anal Prevent* 18, 1-12.

10. Hauer, E. (1992), Empirical bayes approach to the estimation of 'unsafety:' the multivariate regression method, *Accident Anal Prevent* 24, 457-477.
11. Persaud, B.N. (1995), Selection of road sections for safety improvement-comparison of empirical Bayes approaches with More conventional approaches. Working paper, R&D Branch, Ministry of Transportation of Ontario, Toronto, Canada.
12. Persaud, B.N. (1993), Accident potential models for Ontario road sections. report submitted to R&D Branch, Ministry of Transportation of Ontario, Toronto, Canada.
13. Thompson, R.D., E. Lee., and R.M. Thrall (1992), DEA/AR-Efficiency of US independent oil/gas producers over time, *Computer Operations Research* 19, 377-391.
14. Troxel, L. (1993), Severity models for roadside objects. *Transportation Research Circ.* 416, 58-68.

This chapter is based upon Cook, W.D., A. Kazakov and B.N. Persaud (2001). Prioritizing Highway Accident Sites: A Data Envelopment Analysis. Journal of the Operational Research Society, 52, 303-309, with permission from Palgrave.

Chapter 4

BENCHMARKING MODELS

Evaluating the Effect of E-business Activities

4.1. INTRODUCTION

Performance evaluation and benchmarking has become an important continuous improvement tool for business units in the high-technology world of computers and telecommunications where competition is intense and grows more so each day. Benchmarking activities positively force any business unit to constantly evolve and improve in order to survive and prosper in a business environment facing global competition. In fact, as reported in a recent *Wall Street Journal* poll (Lancaster, 1998), benchmarking is one of the top three important and popular tools for continuous improvement. Gap analysis is often used as a fundamental method in performance evaluation and benchmarking. However, as pointed out by Camp (1995), one of the dilemmas that we face is how to show benchmarks where multiple measurements exist. It is rare that one single measure can suffice for the purpose of performance evaluation. As a result, some multi-factor based gap analysis methods have been developed. e.g., Spider charts, AHP maturity index, and Z charts. Although gaps can be identified within each performance measure, it remains a challenging task to combine the multiple measures in the final stage.

Therefore, benchmarking models that can deal with multiple performance measures and provide an integrated benchmarking measure are needed. Note that DEA has been proven an effective tool for evaluating the relative efficiency of peer DMUs when multiple performance measures are present. DEA identifies an efficient frontier (tradeoff curve) along with efficiency scores for all DMUs. Benchmarking is a process of defining valid measures of performance comparison among peer DMUs, using them to determine the relative positions of the peer DMUs and, ultimately, establishing a standard of excellence. In that sense, DEA can be regarded as a benchmarking tool, because the frontier identified can be regarded as an empirical standard of excellence. However, when a new DMU outperforms the identified efficient frontier, a new frontier is generated by DEA. As a result, we do not have the same benchmark for other new DMUs. i.e., the original DEA method needs to be modified as a multi-criteria performance benchmarking tool.

The current chapter presents two DEA-based benchmarking models where the identified efficient frontier (benchmark) remains the same during the benchmarking process. One is called the variable-benchmark model where each DMU under benchmarking is allowed to choose a portion of the benchmark frontier so that the benchmarking performance of the DMU is characterized in the most favorable light. The other is called the fixed-benchmark model where each DMU is benchmarked against the fixed components from the benchmark frontier. The two DEA-based benchmarking are applied to a large Canadian bank (thereafter, CBANK) in measuring the effectiveness of the service delivery configuration.

There are many DEA studies on banking performance. For example, Sherman and Gold (1985) published the first significant DEA bank analysis and started what turned out to be a long list of DEA applications to banking from several different angles (Paradi, Vela and Yang, 2004). Sherman and Ladino (1995) reported that a use of DEA in the restructuring process of the 33 branches of a U.S. bank led to an annual savings of over $6 million. Oral and Yolalan (1990) introduced a DEA model that forced each of 20 branches in a sample to compare itself with the global leader – a Turkish bank.

According to the Canadian Bankers Association, the Canadian banking industry includes 16 domestic banks, 31 foreign bank subsidiaries and 21 foreign bank branches operating in Canada. In total, these institutions manage over $1.7 trillion in assets. Technology innovation is the most important factor contributing to the dramatic changes taking place in Canada's financial services marketplace. Canada's bank financial groups have led the way in providing Canadians with many new ways to access financial services. Canadians have embraced debit cards, ABMs, telephone banking, the Internet and hand-held wireless devices. Advances in technology continue to revolutionize the industry, breaking down geographic barriers and permitting customers to access financial services from virtually anywhere, at any time. In recent years, Canada’s banks have demonstrated a consistent performance, with profits rising significantly from 1995 to 1997.

The rest of the chapter is organized as follows. The next two sections introduce our DEA-based benchmarking models developed in Zhu (2002) and Cook, Seiford and Zhu (2004). We demonstrate how to use the software for DEA benchmarking models. The models are then applied to benchmark a set of e-branches against the best-practice of traditional branches. Concluding remarks are given in the last section.

4.2. VARIABLE-BENCHMARK MODEL

Let E^* represent the set of benchmarks or the best-practice identified by DEA. Based upon the input-oriented CRS envelopment model, we have

$$\begin{aligned} &\min \delta^{CRS} \\ &\text{subject to} \\ &\sum_{j \in E^*} \lambda_j x_{ij} \le \delta^{CRS} x_i^{new} \\ &\sum_{j \in E^*} \lambda_j y_{rj} \ge y_r^{new} \\ &\lambda_j \ge 0, j \in E^* \end{aligned} \tag{4.1}$$

where a new observation is represented by DMU^{new} with inputs x_i^{new} (i = 1, ..., m) and outputs y_r^{new} (r = 1, ..., s). The superscript of CRS indicates that the benchmark frontier composed by benchmark DMUs in set E^* exhibits CRS.

Model (4.1) measures the performance of DMU^{new} with respect to benchmark DMUs in set E^* when outputs are fixed at their current levels. Similarly, based upon the output-oriented CRS envelopment model, we can have a model that measures the performance of DMU^{new} in terms of outputs when inputs are fixed at their current levels.

$$\begin{aligned} &\max \tau^{CRS} \\ &\text{subject to} \\ &\sum_{j \in E^*} \lambda_j x_{ij} \le x_i^{new} \\ &\sum_{j \in E^*} \lambda_j y_{rj} \ge \tau^{CRS} y_r^{new} \\ &\lambda_j \ge 0, j \in E^* \end{aligned} \tag{4.2}$$

Theorem 4.1 $\delta^{CRS*} = 1/\tau^{CRS*}$, where δ^{CRS*} is the optimal value to model (4.1) and τ_o^{CRS*} is the optimal value to model (4.2).

[Proof]: Suppose λ_j^* ($j \in E^*$) is an optimal solution associated with δ^{CRS*} in model (4.1). Now, let $\tau^{CRS*} = 1/\delta^{CRS*}$, and $\lambda_j' = \lambda_j^* \delta_o^{CRS*}$. Then τ^{CRS*} and λ_j' are optimal in model (4.2). Thus, $\delta^{CRS*} = 1/\tau^{CRS*}$. ■

Model (4.1) or (4.2) yields a benchmark for DMU^{new}. The ith input and the rth output for the benchmark can be expressed as

$$\begin{cases} \sum_{j \in E^*} \lambda_j^* x_{ij} & (ith \quad input) \\ \sum_{j \in E^*} \lambda_j^* y_{ij} & (rth \quad output) \end{cases} \tag{4.3}$$

Note also that although the DMUs associated with set E^* are given, the resulting benchmark may be different for each new DMU under evaluation. Because for each new DMU under evaluation, (4.3) may represent a different combination of DMUs associated with set E^*. Thus, models (4.1) and (4.2) represent a variable-benchmark scenario.

Theorem 4.2
(i) $\delta^{CRS*} < 1$ or $\tau^{CRS*} > 1$ indicates that the performance of DMU_o^{new} is dominated by the benchmark in (4.3).
(ii) $\delta^{CRS*} = 1$ or $\tau^{CRS*} = 1$ indicates that DMU^{new} achieved the same performance level of the benchmark in (4.3).
(iii) $\delta^{CRS*} > 1$ or $\tau^{CRS*} < 1$ indicates that input savings or output surpluses exist in DMU_o^{new} when compared to the benchmark in (4.3).

[Proof]: (i) and (ii) are obvious results in terms of DEA efficiency concept.

Now, $\delta^{CRS*} > 1$ indicates that DMU^{new} can increase its inputs to reach the benchmark. This in turn indicates that δ^{CRS*} - 1 measures the input saving achieved by DMU^{new}. Similarly, $\tau^{CRS*} < 1$ indicates that DMU^{new} can decrease its outputs to reach the benchmark. This in turn indicates that 1 - τ^{CRS*} measures the output surplus achieved by DMU^{new}. ■

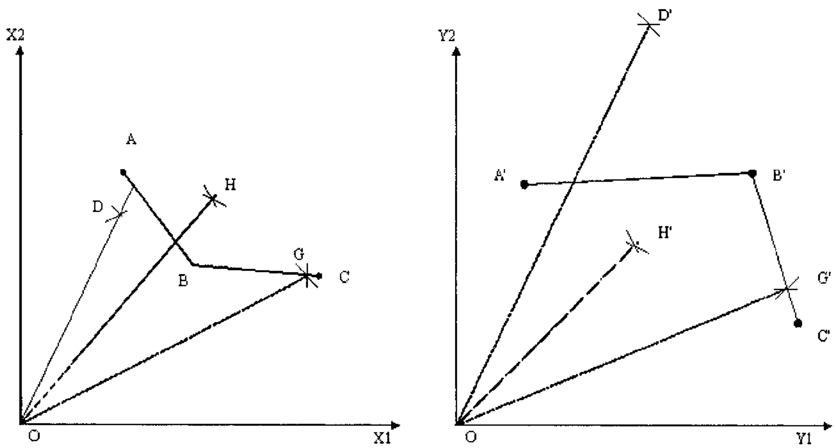

Figure 4-1. Variable-benchmark Model

Figure 4-1 illustrates the three cases described in Theorem 4.2. ABC (A'B'C') represents the input (output) benchmark frontier. D, H and G (or D', H', and G') represent the new DMUs to be benchmarked against ABC (or A'B'C'). We have δ_D^{CRS*} > 1 for DMU D ($\tau_{D'}^{CRS*}$ < 1 for DMU D') indicating that DMU D can increase its input values by δ_D^{CRS*} while producing the same amount of outputs generated by the benchmark (DMU D' can decrease its output levels while using the same amount of input levels consumed by the benchmark). Thus, δ_D^{CRS*} > 1 is a measure of input savings achieved by DMU D and $\tau_{D'}^{CRS*}$ < 1 is a measure of output surpluses achieved by DMU D'.

For DMU G and DMU G', we have δ_G^{CRS*} = 1 and $\tau_{G'}^{CRS*}$ = 1 indicating that they achieve the same performance level of the benchmark and no input savings or output surpluses exist. For DMU H and DMU H', we have δ_H^{CRS*} < 1 and $\tau_{H'}^{CRS*}$ > 1 indicating that inefficiency exists in the performance of these two DMUs.

Note that for example, in Figure 4-1, a convex combination of DMU A and DMU B is used as the benchmark for DMU D while a convex combination of DMU B and DMU C is used as the benchmark for DMU G. Thus, models (4.1) and (4.2) are called variable-benchmark models.

From Theorem 4.2, we can define δ^{CRS*} - 1 or 1 - τ^{CRS*} as the performance gap between DMU^{new} and the benchmark. Based upon δ^{CRS*} or τ^{CRS*}, a ranking of the benchmarking performance can be obtained.

It is likely that scale inefficiency may be allowed in the benchmarking. We therefore modify models (4.1) and (4.2) to incorporate scale inefficiency by assuming VRS.

$$\begin{aligned}
&\min \delta^{VRS} \\
&\text{subject to} \\
&\sum_{j\in E^*} \lambda_j x_{ij} \le \delta^{VRS} x_i^{new} \\
&\sum_{j\in E^*} \lambda_j y_{rj} \ge y_r^{new} \\
&\sum_{j\in E^*} \lambda_j = 1 \\
&\lambda_j \ge 0, j \in E^*
\end{aligned} \tag{4.4}$$

$$\begin{aligned}
&\max \tau^{VRS} \\
&\text{subject to} \\
&\sum_{j\in E^*} \lambda_j x_{ij} \le x_i^{new} \\
&\sum_{j\in E^*} \lambda_j y_{rj} \ge \tau^{VRS} y_r^{new} \\
&\sum_{j\in E^*} \lambda_j = 1 \\
&\lambda_j \ge 0, j \in E^*
\end{aligned} \tag{4.5}$$

Similar to Theorem 4.2, we have

Theorem 4.3
(i) $\delta^{VRS*} < 1$ or $\tau^{VRS*} > 1$ indicates that the performance of DMU^{new} is dominated by the benchmark in (4.3).
(ii) $\delta^{VRS*} = 1$ or $\tau^{VRS*} = 1$ indicates that DMU^{new} achieves the same performance level of the benchmark in (4.3).
(iii) $\delta^{VRS*} > 1$ or $\tau^{VRS*} < 1$ indicates that input savings or output surpluses exist in DMU^{new} when compared to the benchmark in (4.3).

Note that model (4.2) is always feasible, and model (4.1) is infeasible only if certain patterns of zero data are present (Zhu 1996). Thus, if we assume that all the data are positive, (4.1) is always feasible. However, unlike models (4.1) and (4.2), models (4.4) and (4.5) may be infeasible.

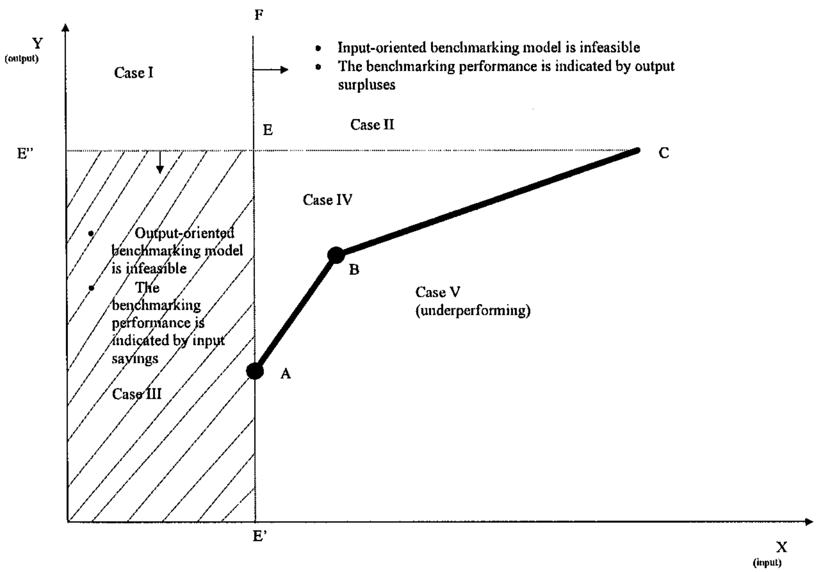

Figure 4-2. Infeasibility of VRS Variable-benchmark Model

Theorem 4.4
(i) If model (4.4) is infeasible, then the output vector of DMU^{new} dominates the output vector of the benchmark in (4.3).
(ii) If model (4.5) is infeasible, then the input vector of DMU^{new} dominates the input vector of the benchmark in (4.3).

[Proof]: The proof follows directly from the necessary and sufficient conditions for infeasibility in the super-efficiency DEA model provided in Seiford and Zhu (1999). ■

The implication of the infeasibility associated with models (4.4) and (4.5) needs to be carefully examined. Consider Figure 4-2 where ABC represents the benchmark frontier. Models (4.4) and (4.5) yield finite optimal values for any DMU^{new} located below EC and to the right of EA. Model (4.4) is infeasible for DMU^{new} located above ray E"C and model (4.5) is infeasible for DMU^{new} located to the left of ray E'E.

Both models (4.4) and (4.5) are infeasible for DMU^{new} located above E"E and to the left of ray EF. Note that if DMU^{new} is located above E"C, its output value is greater than the output value of any convex combinations of A, B and C.

Note also that if DMU^{new} is located to the left of E'F, its input value is less than the input value of any convex combinations of A, B and C.

Based upon Theorem 4.4 and Figure 4-2, we have four cases:

Case I: When both models (4.4) and (4.5) are infeasible, this indicates that DMU^{new} has the smallest input level and the largest output level compared to the benchmark. Thus, both input savings and output surpluses exist in DMU^{new}.

Case II: When model (4.4) is infeasible and model (4.5) is feasible, the infeasibility of model (4.4) is caused by the fact that DMU^{new} has the largest output level compared to the benchmark. Thus, we use model (4.5) to characterize the output surpluses.

Case III: When model (4.5) is infeasible and model (4.4) is feasible, the infeasibility of model (4.5) is caused by the fact that DMU^{new} has the smallest input level compared to the benchmark. Thus, we use model (4.4) to characterize the input savings.

Case IV: When both models (4.4) and (4.5) are feasible, we use both of them to determine whether input savings and output surpluses exist.

If we change the constraint $\sum \lambda_j = 1$ to $\sum \lambda_j \leq 1$ and $\sum \lambda_j \geq 1$, then we obtain the NIRS and NDRS variable-benchmark models, respectively. Infeasibility may be associated with these two types of RTS frontiers, and we should apply the four cases discussed above. Table 4-1 summarizes the variable-benchmark models.

Table 4-1. Variable-benchmark Models

Frontier Type	Input-Oriented	Output-Oriented
CRS	$\min \delta^{Frontier}$ subject to $\sum_{j \in E^*} \lambda_j x_{ij} \le \delta^{Frontier} x_i^{new}$ $\sum_{j \in E^*} \lambda_j y_{rj} \ge y_r^{new}$ $\lambda_j \ge 0, j \in E^*$	$\max \tau^{Frontier}$ subject to $\sum_{j \in E^*} \lambda_j x_{ij} \le x_i^{new}$ $\sum_{j \in E^*} \lambda_j y_{rj} \ge \tau^{Frontier} y_r^{new}$ $\lambda_j \ge 0, j \in E^*$
VRS	Add $\sum \lambda_j = 1$	
NIRS	Add $\sum \lambda_j \le 1$	
NDRS	Add $\sum \lambda_j \ge 1$	

4.3. FIXED-BENCHMARK MODEL

Although the benchmark frontier is given in the variable-benchmark models, a DMU^{new} under benchmarking has the freedom to choose a subset of benchmarks so that the performance of DMU^{new} can be characterized in the most favorable light. Situations when the same benchmark should be fixed are likely to occur. For example, the management may indicate that DMUs A and B in Figure 4-1 should be used as the fixed benchmark. i.e., DMU C in Figure 4-1 may not be used in constructing the benchmark.

To address this situation, we turn to the multiplier models. For example, the input-oriented CRS multiplier model determines a set of referent best-practice DMUs represented by a set of binding constraints in optimality. Let set $\boldsymbol{B} = \{ DMU_j : j \in \mathbf{I}_B \}$ be the selected subset of benchmark set E^*. i.e., $\mathbf{I}_B \subset E^*$. Based upon the input-oriented CRS multiplier model, we have

$$
\begin{aligned}
&\tilde{\sigma}^{CRS*} = \max \sum_{r=1}^{s} \mu_r y_r^{new} \\
&\text{subject to} \\
&\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{s} v_i x_{ij} = 0 \quad j \in \mathbf{I}_B \\
&\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{s} v_i x_{ij} \le 0 \quad j \notin \mathbf{I}_B \\
&\sum_{i=1}^{m} v_i x_i^{new} = 1 \\
&\mu_r, v_i \ge 0.
\end{aligned}
\tag{4.6}
$$

By applying equalities in the constraints associated with benchmark DMUs, model (4.6) measures DMU^{new}'s performance against the benchmark constructed by set $\boldsymbol{B}$. At optimality, some DMU_j $j \notin \mathbf{I}_B$, may join the fixed-benchmark set if the associated constraints are binding.

Note that model (4.6) may be infeasible. For example, the DMUs in set $\boldsymbol{B}$ may not fit into the same facet when they number greater than m+s-1, where

m is the number of inputs and s is the number of outputs. In this case, we need to adjust the set $\boldsymbol{B}$.

Three possible cases are associated with model (4.6): $\tilde{\sigma}^{CRS*} > 1$ indicating that DMU^{new} outperforms the benchmark; $\tilde{\sigma}^{CRS*} = 1$ indicating that DMU^{new} achieves the same performance level of the benchmark; $\tilde{\sigma}^{CRS*} < 1$ indicating that the benchmark outperforms DMU^{new}.

By applying RTS frontier type and model orientation, we obtain the fixed-benchmark models in Table 7-2

Table 4-2. Fixed-benchmark Models

Frontier Type	Input-Oriented	Output-Oriented
	$\max \sum_{r=1}^{s} \mu_r y_r^{new} + \mu$ subject to $\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{s} v_i x_{ij} + \mu = 0 \quad j \in \mathbf{I}_B$ $\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{s} v_i x_{ij} + \mu \le 0 \quad j \notin \mathbf{I}_B$ $\sum_{i=1}^{m} v_i x_i^{new} = 1$ $\mu_r, v_i \ge 0$	$\min \sum_{i=1}^{m} v_i x_i^{new} + v$ subject to $\sum_{i=1}^{s} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} + v = 0 \quad j \in \mathbf{I}_B$ $\sum_{i=1}^{s} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} + v \ge 0 \quad j \notin \mathbf{I}_B$ $\sum_{r=1}^{s} \mu_r y_r^{new} = 1$ $\mu_r, v_i \ge 0$
CRS	where $\mu = 0$	where $v = 0$
VRS	where μ free	where v free
NIRS	where $\mu \le 0$	where $v \ge 0$
NDRS	where $\mu \ge 0$	where $v \le 0$

DMU^{new} is not included in the constraints of $\sum_{r=1}^{s} \mu_r y_{rj}$ - $\sum_{i=1}^{m} v_i x_{ij}$ + $\mu \le 0$ ($j \notin \mathbf{I}_B$) ($\sum_{i=1}^{m} v_i x_{ij}$ - $\sum_{r=1}^{s} \mu_r y_{rj}$ + $v \ge 0$ ($j \notin \mathbf{I}_B$)). However, other peer DMUs (($j \notin \mathbf{I}_B$) are included.

The above models are used in Zhu (2001) in measuring quality of life and Zhu (2004) in evaluating purchasing bids.

4.4. BENCHMARKING MODELS IN DEAFRONTIER SOFTWARE

Zhu (2002) describes how these benchmarking models can be solved in Microsoft® Excel and Excel Solver. Here, we demonstrate how these benchmarking models can be solved using the DEAFrontier software.

To run the variable-benchmark models presented in Table 4.1, we need to set up the data sheets. *Store the benchmarks in a sheet named "Benchmarks" and the DMUs under evaluation in a sheet named "DMUs"*. The format for these two sheets is the same as that shown in Figure 1-10. Then select the

Variable Benchmark Model menu item. You will be prompted with a form for selecting the model orientation and the frontier type as shown in Figure 4-3. Note that if you select a frontier type other than CRS, the results may be infeasible. The benchmarking results are reported in the sheet "Benchmarking Results".

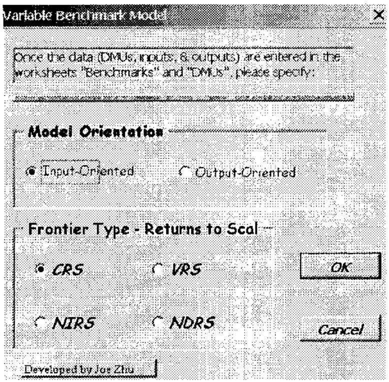

Figure 4-3. Variable Benchmark Models in DEAFrontier

To run the fixed-benchmark models presented in Table 4-3, we *store the benchmarks in a sheet named "Benchmarks" and the DMUs under evaluation in a sheet named "DMUs"*. Then select the Fixed-Benchmark Model menu item. You will be prompted with a form for selecting the model orientation and the frontier type. The results are reported in the "Efficiency Report" sheet. If the benchmarks are not properly selected, you will have infeasible results and need to adjust the benchmarks.

4.5. APPLICATION TO BANK BRANCHES

In the financial services industry worldwide, the traditional face-to-face customer contacts are being replaced by electronic points of contact to reduce the time and cost of processing an application for various products. To best respond to this new marketplace, the CBANK identified a need to conduct research into the design and delivery of financial services by the most efficient and effective means while meeting internal operational performance goals. CBANK created a set of 12 e-business branches (thereafter, e-branches[1]) using a new banking concept intended to create customer convenience with more efficient platforms for performing

[1] We code these e-branches as e1, e2, …, e12.

transactions. The e-branches are aimed at increasing the speed of service delivery and decreasing costs in significant proportions through branch operation automation via Internet, ATMs, telephone banking and other electronic means. From a business perspective, these e-branches are a result of application of technology toward the automation of business transactions and workflows (Kalakota and Whinston, 1997).

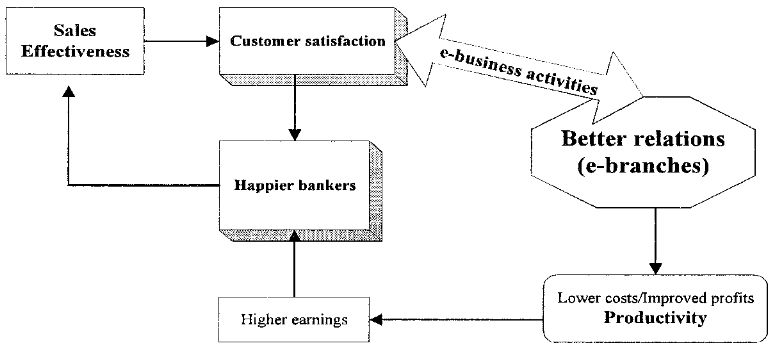

Figure 4-4. The effect of e-business activities on banking performance

Figure 4-4 presents the impact of e-business on banking performance. Based upon Harvey (1996), sales effectiveness/customer satisfaction leads to better relationships with a bank branch's current clients who are inclined to bring more of their business to it. This increases market share, as does the influx of new customers who hear about the branch's legendary levels of service. Increased share means that more transactions are being processed, presumably with the same amount of fixed cost. That, in turn, lowers unit cost and increases revenues, both of which lead to improved productivity and higher earnings. More earnings result in increases in the bonus pool, higher merit increases, and a higher stock price, which benefits all shareholders and the workforce. This leads to happier banks. We should not forget that the reverse of the cycle is also true. If a bank branch does not provide the level of service that people want, it will lose customers to the competition. Since the e-branches are in a new form of business and banking is done in very different ways, uncertainty surrounds the development of this new delivery model.

Performing a benchmarking study is extremely critical to the CBANK in undertaking the e-business activities and to examine whether the e-branches enhance the productivity and sales effectiveness while reducing expenditures. i.e., the CBANK wants to examine whether the e-business activities have a positive impact on the business cycle presented in Figure 3. Clearly productivity in this case is characterized by a number of measures including, labor, information technology investment, transactions, and others.

We develop two DEA inputs: FTE (full time equivalent) counts which include sales, service, support and other staff, and operating expenses which include spending on stationary, communications, shortage & losses, business development, employee training, advertising & publications, computer costs, and others.

Table 4-3. Transactions and Processing Time

		Processing Time (hour)	
DEA Outputs	Transaction Description	Min.	Max.
1 (Tran1)	Set up a new collateral or non collateral loan	.798	2.093
2 (Tran2)	Open a new account (menu account)	.353	.541
3 (Tran3)	Process a branch deposit to menu account	.044	.070
4 (Tran4)	Process a withdrawal from menu account	.040	.055
5 (Tran5)	Update passbook for menu account in branch	.018	.026
6 (Tran6)	Process visa cash advance	.041	.101
7 (Tran7)	Process a business deposit	.039	.092

There are two types of transactions: sales and service. We select seven transactions (DEA outputs) presented in Table 4-3 that account for over 90% of the volume of sales and service related work carried out.

4.5.1 Identification of Benchmark frontier

This section identifies the best practice frontier of traditional bank branches in each quarter from 1995 to 1996[2]. The identified best-practice in each quarter is later used as the benchmark frontier to evaluate the quarterly performance of e-branches.

Table 4-3 reports the minimum and maximum process times for the seven transactions. The minimum and maximum process times are used as lower and upper bounds for output multipliers in the multiplier models to develop the following weight restrictions.

$$1.475 \le \frac{Tran_1}{Tran_2} \le 5.929,\ 11.271 \le \frac{Tran_1}{Tran_3} \le 47.568,$$

$$14.345 \le \frac{Tran_1}{Tran_4} \le 52.325, 30.346 \le \frac{Tran_1}{Tran_5} \le 116.278,$$

$$7.812 \le \frac{Tran_1}{Tran_6} \le 51.049, 8.576 \le \frac{Tran_1}{Tran_7} \le 53.667$$

There are about 1200 traditional branches within the CBANK. We will identify the efficient ones in each quarter to use as a benchmark data set. Both CRS and VRS multiplier models in Table 1-2 with the above AR restrictions are applied to identify the quarterly best-practice frontier.

4.5.2 Benchmarking the e-branches against the traditional branches

We benchmark the 12 branches (e-branches) against the identified best-practice of traditional branches in each quarter from 1995 to 1996. The last quarter of 1996 is regarded as the "turning point", since the e-branches were created during the last quarter of 1996. Note that the best-practice of traditional branches was changing quarterly. Thus, we here capture a dynamic picture of the performance change of these 12 e-branches.

The results from the CRS variable-benchmark model (4.1) indicate a dramatic performance change from the third quarter to the fourth quarter in 1995 when the automation is implemented: all the e-branches outperform the best-practice of traditional branches. However, the performance of the e-branches decline into 1996. Based upon the optimal value to model (4.1) we classify the e-branches into four categories with respect to the performance change during 1996 (see Figure 4-5):

[2] The current study uses the real data from 1995 to 1996, since the CBANK had just started the service re-designing for the 12 branches. After that period, the CBANK had stopped the re-designing of these branches.

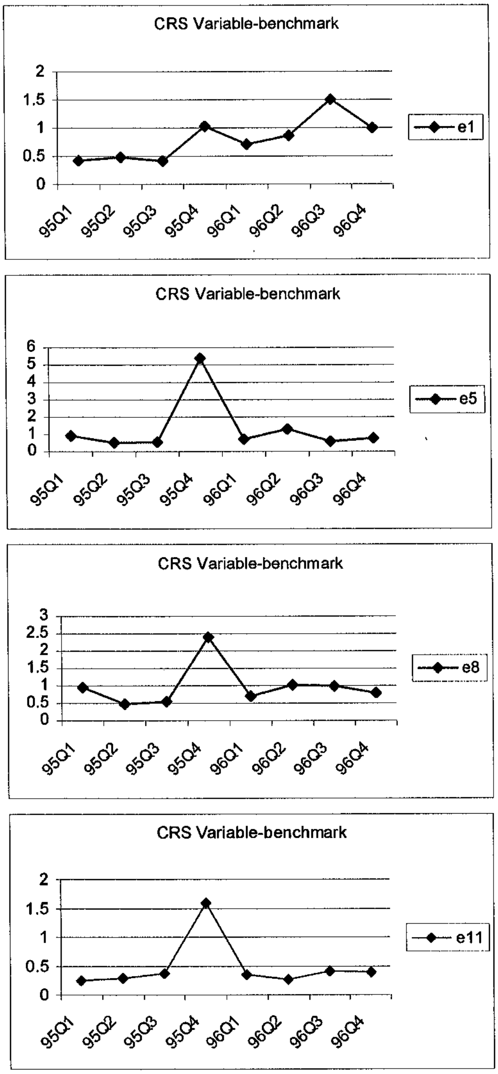

Figure 4-5. Representative performance change patterns when the e-branches are benchmarked against the traditional branches

(i) branches, e1, e3, and e6 where the performance improves over the first three quarters of 1996 and then declines;
(ii) branches e2, e5, e7, and e12 where the performance improves from 96Q1 to 96Q2, then declines from 96Q2 to 96Q3, and then improves from 96Q3 to 96Q4[3];
(iii) branches e4, e8, e9 and e10 where performance improves from 96Q1 to 96Q2 and then declines;
(iv) branches e11 where the performance declines from 96Q1 to 96Q2, and then improves.

Only 5 e-branches, in the second quarter of 1996, and 3 e-branches in the third quarter of 1996 outperformed the best-practice. i.e., the optimal value to model (4.1) – δ_o^{CRS*} is greater than one. The majority of the e-branches did not show performance improvement compared to the best-practice of traditional branches.

Recall that model (4.1) assumes CRS, i.e., scale inefficiency is not allowed. We therefore turn to VRS models. Similar results are obtained. This indicates that scale is not a factor related to the productivity in the e-branches.

Under the case of VRS (models (4.4) and (4.5)), no infeasibility occurs, because most of the e-branches are under-performing units compared to the benchmark, this indicates that most of the e-branches are of Case V in Figure 4-2.

4.5.3 Benchmarking within e-branches

The previous analysis indicates that there was no productivity gain as a result of branch automation when the performance of e-branches is compared to the best-practice of traditional branches. Recall that the 12 e-branches were re-engineered from 12 existing traditional branches which were under-performing units compared to the best-practice in the first three quarters in 1995. Thus, it might be difficult for the newly established e-branches to close the performance gap between the best-practice and their predecessors. Therefore, we next study the performance change within the 12 branches. i.e., we compare the e-branches to the best-practice of these 12 branches before the automation.

First, we compare the e-branches in each quarter of 1996 to the best-practice of e-branches in each quarter of 1995[4]. The results from models (4.1), (4.4) and (4.5) show that (i) when the best-practice of the e-branch

[3] 96Q1 stands for the first quarter of 1996.
[4] The "e-branches" in 1995 are referred to the 12 branches before the automation.

predecessors in the first quarter of 1995 is used as the benchmark, most of the e-branches outperform the benchmark, although the performance of most e-branches declines into the last quarter of 1996, and (ii) under the assumption of VRS, two e-branch moved from Case II into Case IV described in Figure 4-2, indicating a productivity decline and most of the e-branches can be categorized by Case III in Figure 4-2. Only one e-branch in the second quarter of 1996 is of Case I, representing the best scenario with lowest costs and highest performance.

Overall, the performance of e-branches declines as the benchmark is changed from the first quarter to the last quarter of 1995[5].

Next, we assume each branch in each quarter of 1995 represents a branch. Since automation happened during the last quarter of 1995, we exclude the branches in that quarter from the identification of best-practice. Thus, we have 12 (branches) × 3 (quarters in 1995) = 36 branches. We then benchmark the e-branches in the last quarter of 1995 and in each quarter of 1996 against the best practice of these 36 branches.

In these 36 branches, eight branches, namely, e1-95Q1, e1-95Q2, e3-95Q1, e5-95Q2, e6-95Q2, e5-95Q3, e6-95Q3, and e10-95Q3, are best-practice branches. These eight branches are used as benchmarks in model (4). In model (9), we select e5-95Q3, e695Q3, and e10-95Q3 as three fixed benchmarks, since they represent the best-practice right before the establishment of e-branches.

Tables 4-4 and 4-5 report the benchmarking scores from models (4.1) and (4.6) respectively. These are optimal values to models (4) and (9). A larger value indicates a better performance. For example, under 96Q1 of Table 4-4, e1 has a score of 2.1658 when e1 is compared to the best-practice of the traditional branches in the first quarter of 1996. This indicates that e1 in the first quarter of 1996 outperformed the traditional branches. For e2, the corresponding benchmarking score is 0.3942, indicating that e2 was dominated by the traditional branches.

Overall, the performance of these e-branches declines from the first quarter to the last quarter of 1996.

When the variable-benchmark model is used, the benchmarking performance of e3, e6 and e9 constantly declines. The performance of e1

[5] However, as pointed by one reviewer, since the conversion to e-branches took place during the last quarter of 1995, this particular quarter may be tainted with the effects of the conversion. Also, not all the branches are converted in the same day. As a result, quarter totals include a mix of new and traditional branches. On the other hand, it is possible that these branches were already producing at a higher level during 1995, and therefore there were no noticeable changes once e-branch conversion took over. The one quarter performance jump can be due to the novelty aspect as customers would come and check things out.

which outperforms the best-practice improves during the first three quarters and then declines. Note that only one branch (e11)'s performance declines, outperform the best-practice, and then declines. The remaining 7 branches show a performance improvement from the first to the second quarter of 1996 and then show a constant performance decline with respect to the best-practice.

Table 4-4. Benchmarking within e-branch: variable-benchmark model

e-branches	95Q4	96Q1	96Q2	96Q3	96Q4
e1	2.3770	2.1658	2.3860	3.3878	1.0513
e2		0.3942	2.0119	0.3632	0.2353
e3	16.2577	0.7857	0.7101	0.6768	0.1803
e4	8.6272	1.1054	1.1725	0.8412	0.1988
e5	32.4540	0.9147	1.1725	0.7577	0.3429
e6	12.9585	1.5031	1.4502	1.4078	0.4256
e7	6.1554	0.4223	0.6771	0.3753	0.1488
e8	14.1309	0.8519	0.9365	0.7965	0.2839
e9	11.5337	1.1467	1.0672	0.8693	0.2227
e10	7.6262	0.9750	1.7775	0.8508	0.3504
e11	9.0184	1.2259	0.5872	1.0533	0.2725
e12	8.4983	0.9209	1.7458	1.0977	0.4279

Table 4-5. Benchmarking within e-branch: fixed-benchmark model

e-branches	95Q4	96Q1	96Q2	96Q3	96Q4
e1	2.2038	1.4076	1.2771	1.4672	0.4706
e2		0.2818	0.8540	0.3332	0.1826
e3	10.9393	0.7857	0.6931	0.6516	0.1764
e4	6.6742	1.1054	1.0399	0.7302	0.1356
e5	21.8097	0.8787	1.0496	0.7499	0.3406
e6	9.3977	1.0599	1.0730	1.0539	0.3274
e7	4.0593	0.4152	0.4761	0.3746	0.1460
e8	9.5579	0.7892	0.7187	0.7094	0.2567
e9	7.8706	1.0149	1.0488	0.8693	0.2133
e10	5.8255	0.9750	1.3070	0.8508	0.3092
e11	6.3961	1.2092	0.5789	0.7750	0.1960
e12	6.5173	0.9209	1.2823	1.0977	0.3856

When the fixed-benchmark model is used, as expected the benchmarking scores decrease, implying a worse performance with respect to the best-practice. The performance change of e1 and e11 remains the same patterns as those under the variable-benchmark model. The benchmarking performance of e3, e4, e6 and e8 constantly declines. The remaining 6 branches show a performance improvement from the first to the second quarter of 1996 and then show a constant performance decline with respect to the best-practice.

Finally, we note that an additional factor may contribute to the dramatic performance improvement when the CBANK launched the e-branches in the last quarter of 1995. There were some additional staff employed by the e-branches that were not reported as they were under the balance sheet of the head office and not paid for by the e-branches. The CBANK phased out these unreported staff over the first two periods (the last quarter of 1995 and the first quarter of 1996). Thus, the number of staff in the last quarter of 1995 and the first quarter of 1996 was understated. As a result, the resulting benchmarking scores should be decreased in the last quarter of 1995 and the first quarter of 1996. Such adjustment indicates that the performance change of e-branches does not move in a favorable direction.

The above analysis indicates that the e-business activities (establishing the e-branches) do not lead to an increased productivity. This empirical finding helps the CBANK to further examine its current e-business options.

4.6. CONCLUSIONS

To aid the CBANK in undertaking e-business activities, the current study is directed at evaluating and benchmarking branch bank performance. Two DEA-based benchmarking models are developed to study the change in performance that branches undergo when moving from the old to the new structure where transactions are automated. The study reveals that e-branches (new structure) did not exhibit productivity gain when compared to both the best-practice of traditional branches and e-branches' predecessors. This finding allows the bank to examine its business options, and gain an understanding of what does not work well in terms of the makeup of new branches. This further can point to weaknesses and strengths in e-branch operations. The current study provides tools needed to monitor the performance change and further facilitates the development of the best strategic option for the organization with regard to branch makeup.

REFERENCES

1. Camp, R.C. (1995), Business Process Benchmarking, Finding and Implementing Best Practices. ASQC Quality Press, Milwaukee, Wisconsin.
2. Cook, W.D., L.M. Seiford. and J. Zhu (2004), Models for performance benchmarking: measuring the effect of e-business activities on banking performance, *OMEGA* 32, 313-322.
3. Harvey, T. (1996), The Banking Revolution: positioning your bank in the new financial service marketplace. IRWIN, A Bankline Publication.
4. Kalakota, R. and A.B. Whinston (1997), Electronic Commerce: A Manager's Guide. Addison-Wesley.
5. Lancaster, H. (1998), Managing your career, Wall Street Journal, Tuesday, April 28, B1.
6. Oral, M. and R. Yolalan (1990), An empirical study on measuring operating efficiency and profitability of bank branches, *European Journal of Operational Research* 46, 282-249.
7. Paradi, J.C., S. Vela and Z. Yang (2004), Assessing bank and bank branch performance: modeling considerations and approaches, Chapter 13, 349-400, in Handbook on Data Envelopment Analysis, eds W.W. Cooper, L.M. Seiford and J. Zhu, Kluwer Academic Publishers, Boston.
8. Seiford, L.M. and J. Zhu (1999), Infeasibility of super-efficiency data envelopment analysis models, *INFOR* 37, 174-187.
9. Sherman, H.D. and Gold, F. (1985), Bank branch operating efficiency: evaluation with data envelopment analysis, *Journal of Banking and Finance* 9, 297-316.
10. Sherman, H.D. and G. Ladino (1995), Managing bank productivity using data envelopment analysis (DEA), *Interfaces* 25, 60 – 73.
11. Thompson, R.G., Langemeier, L.N., Lee, C., Lee, E., and Thrall, R.M. (1990), The role of multiplier bounds in efficiency analysis with application to Kansas farming, *Journal of Econometrics* 46, 93-108.
12. Violano, M., and S.V. Collie (1992), Retail Banking Technology. John Wiley & Sons, Inc., New York.
13. Zhu, J. (1996), Robustness of the efficient DMUs in data envelopment analysis, *European Journal of Operational Research* 90 451-460.
14. Zhu, J. (2001), A multidimensional quality-of-life measure with an application to Fortune's best cities, *Socio-Economic Planning Sciences* 35, 263-284.
15. Zhu, J. (2002), Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets and DEA Excel Solver. Kluwer Academic Publishers, Boston.

16. Zhu, J. (2004), A buyer-seller game model for selection and negotiation of purchasing bids: extensions and new models, *European Journal of Operational Research* 154 150-156.

This chapter is based upon W.D. Cook, L.M. Seiford. and J. Zhu, 2004, Models for Performance Benchmarking: Measuring the Effect of E-business Activities on Banking Performance, OMEGA, , with permission from Elsevier Science.

Chapter 5

FACTOR SELECTION ISSUES IN BANK BRANCH PERFORMANCE

5.1. INTRODUCTION

Although many studies and applications have demonstrated the effectiveness of DEA, it remains that for large-scale problems, with many different factors or variables available, at least *two* impediments to effective implementation may still exist. First, it is recognized that a DEA analysis entails explicitly specifying a set of factors to be used in the model, and classified as to which will constitute outputs or results, and which are inputs or impacts. In many settings, however, it can be problematic to define the most appropriate of those factors to be integrated into the analysis (and their status as to input or output); as with conventional statistical analyses, many choices can exist. A *second* major element involves implementation, and has to do with management's own perceptions as to what constitutes good versus poor performance. If a methodology fails to uncover what management feels is best or worst practice, that methodology is unlikely to succeed as the measurement tool of choice. This is particularly the case in those settings such as banks, where established procedures are in place to formally track the activities of the DMUs (e.g., branches).

In this chapter we examine how expert knowledge in the form of classification information can be incorporated within the DEA structure, to enhance performance measurement. It must be said initially that management perceptions as to good versus poor performance may or may not be enlightened. "Expert knowledge" may be nothing more than uninformed opinion or bias, and as such should not be brought to bear on performance evaluation. In many cases, however, expeert opinion is in fact grounded in a solid knowledge of the situation at hand.

Consider the example involving measuring the relative efficiencies of highway maintenance crews as discussed in Cook et al. (1990) and Cook and Zhu (2003). In the early stages of the development of the DEA model there, district geotechnical staff and maintenance supervisors were consulted regarding the work load of maintenance crews in the study area. From those discussions, a choice was made regarding appropriate inputs and outputs.

In the preliminary analysis no provision was made for winter versus summer maintenance, with the result that maintenance crews (patrols) in the 'snow belt' tended to all receive very low efficiency scores. It was management's conviction, however, that a number of those crews were in their words, *efficient*, arguing that the greater effort by those crews in the form of more frequent use of snow removal equipment, and application of road salting, needed to be taken into account. With this, it was realized that an output measure accounting for the winter factor was necessary. When this modification was incorporated, it indeed did occur that certain crews were, under the initial model, being unduly penalized in the absence of a proper definition of factors.

This example serves to demonstrate that management opinion may not necessarily be given in the form of explicit identification of input/output factors, but rather is often expressed in a more global 'sense' of DMUs being efficient versus inefficient. In many circumstances, this form of expression of expertise can be a valuable input to the performance measurement exercise.

Section 5.2 presents the problem setting for a particular DEA analysis undertaken in a major Canadian bank. With this setting as a backdrop, the remainder of the chapter sets out to describe a DEA model augmented by branch classification information. We argue that the latter is a form of expert knowledge that should be accounted for in any DEA analysis for such a performance measurement situation. In Section 5.3 we review various discriminant models that are applied in the aforementioned first stage. Section 5.4 discusses the broad structure of the expert-enhanced DEA model. As discussed above, we view the modeling of performance measurement as a two-stage process. In the *first* stage, a classification or discriminant model is used to designate the status (output or input) for each variable; in the *second*, the DEA analysis is performed, based on the variable designations chosen. In this first set of experiments, *all* variables are assumed to be flexible as to their input versus output status. An extensive simulation experiment is conducted in Section 5.5, using data supplied on 200 bank branches, and classification data provided by branch consultants. Section 5.6 utilizes a particular discriminant model, and adds the feature that certain of the variables are inflexible (i.e. are locked in from the beginning as inputs or as outputs). Outcomes from this resulting second set of experiments are presented.

5.2. THE PROBLEM SETTING

The current chapter documents a case analysis of performance measurement of branches within a major Canadian bank. The distinguishing feature of this application, in comparison to others in a similar setting, is the presence of an existing performance measurement system. As with many banks, this particular one employs branch consultants who closely monitor branch operations by collecting detailed time study information on each process. This leads to estimates of processing times that help consultants to classify branches into various categories as to their efficiency status. This existing procedure for classifying branches is inexact and not transparent. It combines both quantitative information based on time studies, as well as demographic (customer-base) data. Specifically, consultants will attempt to evaluate branches based upon their *potential to perform.* An example demographic parameter would be to percent of high value customers in their customer base. A large percentage of high value customers generally implies a potential for a higher than normal level of product sales, (e.g. mutual funds) in that branch. Such factors are part of the rationale for a final classification.

Attempts to apply conventional DEA principles here met with some implementation difficulties. Essentially, the frontier branches arising from the DEA analysis in many instances do not coincide with the classification identified by branch consultants. In a pure operational efficiency sense, DEA results are an accurate portrayal of branch performance, at least from the perspective of *technical efficiency.* In a pragmatic sense, however, the analysis appears to be failing to capture those elements, many qualitative, that branch consultants take into account.

Table 5-1. DEA Variables

Variable	Description
FTETOT	The sum of all full time employees (sales and service positions)
RSP	The number of retirement savings plans sold
LOANTOT	The total of all loans and mortgages
MOPCAO	The total of accounts opened
MDPMTRF	The number of deposits and transfers
MWDMUPD	The number of withdrawals and updates

Data for six operational variables were supplied for the study, as presented in Table 5-1.

These variables capture the business activity of branches. One difficulty of applying conventional DEA in this situation is defining the (input versus output) status of each of the supplied variables. One can claim that a variable such as total withdrawals and updates should be considered as an *output* in that it is a part of the branch's workload. However, if the strategy of the bank is sales oriented, then this variable may, in fact, create a deterrent to sales generation, thus representing a type of *environmental input*. Other variables such as mortgages and loans are clearly outputs, and no dispute exists as to their status.

This performance measurement setting thus presents the challenge of combining two forms of data–quantitative data on a set of supplied variables, and qualitative data in the form of a classification of branches supplied by bank consultants. Appendix I presents a *sample* selected from a set of 200 branches that have been classified as high or low in terms of performance. Since the current approach to performance measurement (and the qualitative factors that lead to that approach) appears in the form of a classification of the branches, the methodology proposed herein attempts to capture that information as well as the data provided in Table 5-1. As indicated above, we propose a two stage approach to performance measurement. In the first stage the current classification information is used to aid in designating the status (input versus output) of variables to be used in the DEA analysis. In the second stage, the *resulting* inputs and outputs are used to perform a DEA analysis leading to branch performance measures. To facilitate *stage one*, various discriminant tools are examined in the following section.

These variables capture the business activity of branches. One difficulty of applying conventional DEA in this situation is defining the (input versus output) status of each of the supplied variables. Arguably one can claim that a variable such as total withdrawals and updates should be considered as an *output* in that it is a part of the branch's workload. However, if the strategy of the bank is sales oriented, then this variable may, in fact, create a deterrent to sales generation, thus representing a type of *environmental input*. Other variables such as mortgages and loans are clearly outputs, and no dispute exists as to their status.

This performance measurement setting thus presents the challenge of combining two forms of data–quantitative data on a set of supplied variables, and qualitative data in the form of a classification of branches supplied by bank consultants. Appendix I presents a *sample* selected from a set of 200 branches that have been classified as high or low in terms of performance. Since the current approach to performance measurement (and the qualitative factors that lead to that approach) appears in the form of a classification of

the branches, the methodology proposed herein attempts to capture that information as well as the data provided in Table 5-1. As indicated above, we propose a two stage approach to performance measurement. In the first stage the current classification information is used to aid in designating the status (input versus output) of variables to be used in the DEA analysis. In the second stage, the *resulting* inputs and outputs are used to perform a DEA analysis leading to branch performance measures. To facilitate *stage one*, various discriminant tools are examined in the following section.

5.3. DISCRIMINANT MODELS

As indicated above, the aim of this chapter is to demonstrate the use of discriminant tools to provide a link between expert knowledge in the form of classification information, and conventional quantitative data, thereby providing a format for both data sources in the conventional DEA structure. To facilitate the discussion in later sections, we provide here a brief review of some of the standard discriminant models that have been applied to our data. Specifically, we examine logistic regression (LR), multiple discriminant analysis (MDA), goal programming (GP), and integer goal programming (IGP).

5.3.1 Logistic Regression

The logistic regression (LR) technique, Kleinbaum (1994), analyzes the relationship between a categorical dependent (or response) variable, and a set of independent (or explanatory) variables. A principal model within LR is the Logit model, which has only two categories in the response variable – event A or non-A.

To place the problem in a general framework, let R denote the random response variable of interest (in the present setting, the response corresponds to the categorization of n bank branches as high or low performers). For each branch $k = 1,\ldots,n$, let $Z_k = (z_{qk})_{q=1}^{Q}$ be a Q-dimensional vector of variables.

$$E(R_k = 1) = \frac{\exp\left(\sum_{q=1}^{Q} b_q z_{qk}\right)}{\left[1 + \exp\left(\sum_{q=1}^{Q} b_q z_{qk}\right)\right]} = P_k. \quad (5.1)$$

Here, the b_q are the multipliers (regression coefficients) of the Q variables, z_{qk}, and P_k denotes the probability that branch k will be classified as a high performer. The probability of the non-event (branch is a low performer) is then:

$$E(R_k = 0) = 1 - \frac{\exp\left(\sum_{q=1}^{Q} b_q z_{qk}\right)}{\left[1 + \exp\left(\sum_{q=1}^{Q} b_q z_{qk}\right)\right]} = 1 - P_k.$$

Therefore, we can state that

$$E(R_k = 1) / E(R_k = 0) = P_k / (1 - P_k) = \exp\left(\sum_{q=1}^{Q} b_q z_{qk}\right).$$

The fraction $P_k / (1 - P_k)$ is called the *odds ratio.* Now, take the natural log of the odds ratio

$$L = L_n\left(\frac{P_k}{1 - P_k}\right) = \sum_{q=1}^{Q} b_q z_{qk},$$

with $\lim_{\beta' z \to +\infty} \text{Prob}(\text{R}_\text{k} = 1) = 1$ and $\lim_{\beta' z \to +\infty} \text{Prob}(\text{R}_\text{k} = 1) = 0$. L is called the *logit*, and hence the name *logit model*. Here β_q denotes the population regression coefficient of which b_q is an estimate.

5.3.2 Multiple Discriminant Analysis

The basic purpose of multiple discriminant analysis (MDA) is to estimate the relationship between a single non-metric (categorical) dependent variable (groups), and a set of metric independent variables (predictors). MDA, which can classify two or more groups, identifies the areas where the greatest difference exists between the groups, derives a discriminant weighting coefficient for each variable to reflect these differences, and then assigns each individual to a group, using the weights and each individual's ratings on the characteristics. The ultimate goal in MDA is to predict to which group a new observation belongs.

MDA is based on centroids and groups. It involves deriving the linear combination of the two (or more) independent variables that will discriminate best between the a priori defined groups. This is achieved by the statistical decision rule of maximizing the between-group variance, relative to the within-group variance. This relationship is expressed as the ratio of between-group to within-group variance. If the variance between is large relative to the variance within the groups, we say that the discriminant function separates the groups well.

The test for the statistical significance of the discriminant function is a generalized measure of the distance between the groups centroids. This is done by comparing the distribution of the discriminant scores for the two groups. If the overlap in the distribution is small, the discriminant function separates the groups well.

5.3.3 Goal Programming Models

Applications of linear goal programming-based approaches in discriminating between two groups of observations, have appeared in numerous publications, e.g. Mignona and Glover (1995). With GP we seek a hyperplane to separate two groups of points in the best possible way, regardless of whether or not they can be completely separated. See Glover (1990).

Assume that the n bank branches have been grouped into two categories G_1, G_2, where G_1 represents branches classified as low performers, and G_2, those considered as high performers. Freed and Glover (1986) present several variations of a goal programming model for discriminating between two groups. Stated in simple terms, the GP model is

$$\min \sum_{k=1}^{n} k_k \alpha_k$$

subject to:

$$\sum_{q=1}^{Q} b_q z_{qk} - \alpha_k \le T, \quad k \varepsilon G_1,$$

$$\sum_{q=1}^{Q} b_q z_{qk} + \alpha_k \ge T, \quad k \varepsilon G_2, \tag{5.2}$$

$$\alpha_k \ge 0,\ \forall k;\ T, b_q \text{ unrestricted in sign.}$$

Here, the α_k are goal achievement variables, and T is a variable representing the threshold against which branch performance is compared.

The goal is to minimize the weighted sum of deviations. In problems where it is especially important to correctly classify certain observations, those observations can be weighted by increasing the appropriate h_k values in the objective function.

5.3.4 IGP Models

The integer programming formulation, where the objective is to minimize the number of misclassified points, can be stated as follows:

$$\min \sum_{k=1}^{n} \gamma_k$$

subject to

$$\sum_{q=1}^{Q} b_q z_{qk} - M\gamma_k \le T, \quad k \varepsilon G_1,$$

$$\sum_{q=1}^{Q} b_q z_{qk} + M\gamma_k \geq T, \quad k\varepsilon G2,$$

$$\gamma_k \varepsilon \{0,1\}, \quad T, b_q \text{ unrestricted in sign,}$$

where M is a large positive number, and the γ_k are binary variables used to count the number of violations.

5.4. EMBEDDING EXPERT KNOWLEDGE IN THE ADDITIVE DEA MODEL

Consider the situation in which management has provided an expert opinion in the form of a classification of DMUs into two principal groups - call them good and poor performers. We now wish to apply the principles of DEA to derive a measure of performance for each member of an entire set of DMUs, but in a way that embeds this classification information into the model structure. If one were to develop an expert system, an appropriate question to ask here would be 'what functional relationship among the available variables (e.g., sales, staff size, deposits, etc.), would provide a classification of the DMUs that most closely resemble management's classification?' Any expert system works essentially in this way. The assumption must be, of course, that this form of information does in fact constitute expertise, and that we actually do wish to create a model that can come as close as possible to replicating the expert's view of performance.

In the context of DEA, an analogous interpretation of this idea is to pose the question as 'which variables should serve as outputs and which as inputs, such that the DEA analysis produces performance measures that are clustered in a way that best imitates management's classification?' Such a DEA model will then be a form of expert system performance measurement tool. The most basic method for embedding expert opinion into the DEA structure is a two-stage process. In the first stage, a classification model is applied to aid in choosing which variables to designate as outputs and which as inputs. In the second stage, the DEA model is applied to derive a performance measure for each DMU. The hypothesis is that the DEA scores, so derived, will be consistent with management's opinions. Specifically, when ranked, the scores will provide a clustering of the sample DMUs into two groups that imitate the groupings provided by the experts. *In this chapter we set out to test this hypothesis.*

Discriminant techniques are particularly helpful in variable selection in this context as they:

- use the branch consultant's knowledge in terms of branch discrimination; and

- do not depend upon prescribed variable orientations

Some parameters are often not analyzed in a DEA analysis (e.g. environmental data, demography, fixed inputs ...), but may well have been part of management's mental model in classifying the branches. Discriminant techniques can assist in extracting classification knowledge, and use this information to select appropriate variables, by orienting them to produce results generally consistent with management's perceptions. In this way, the resulting DEA model incorporates a broad range of factors, both explicit, and implicit.

Each discriminant technique (see discussion in the subsection to follow) computes a discriminant function $f_i(x_{ij}, c_{ij}) = \gamma_i$, where each variable has a coefficient (or weight). Here, γ_i can be a probability (i.e., logistic regression), or a scalar (i.e., goal programming). The nature of γ_i is not particularly important, in that it is used only as a classification measure, based on a threshold (or cutoff value). This threshold determines if the observation i with the score γ_i belongs to group 1 or 2 (with two groups cases).

The principal objective of the experiment carried out in this chapter is to provide an improved DEA model that utilizes branch consultants' judgment. We reiterate that in our particular case this judgment, or knowledge, is represented by the classification of a set of bank branches into two groups... high and low performers. While the particular problem setting herein classifies branches via expert opinion, the same idea applies in situations where classification can arise in other ways (e.g. bankrupt versus non-bankrupt firms). The basic hypothesis is that the sign of a discriminant function coefficient can be used to determine if the corresponding variable should be considered as an input, or an output in a DEA model. This approach can be very useful when a DEA problem has flexible variables (variables that could be either inputs or outputs).

One can put forward at least three reservations concerning this approach:

1. *Is the current application one in which there is flexibility concerning the status (output or input) of some of the candidate variables*?

In many applications, it is the case that variable status is well defined. There is generally no dispute, for example, as to the (input) status of branch *staff* in bank branch performance analysis. It has, however, been recognized in previous applications (see, e.g., the highway maintenance application of Cook et al.(1991)), that it can be difficult to decide on the status of certain factors. For example, is the road surface condition variable in the highway maintenance problem, an environmental input variable that influences the amount of maintenance resources that need to be applied to the road network, or a discretionary output that reflects the quality of earlier maintenance work done? In a bank setting, should one consider back office work (filing, etc.) as an output, reflecting the work carried out by the branch, or as an input that deters staff from performing perhaps more important duties? We, therefore, contend that there is sufficient scope for flexibility in many applications, including the current one, as to variable status.

2. *Should the branch consultants' knowledge base, pertaining to branch performance, be viewed as a level of expertise that is worthy of incorporation into a measurement model?*

There is no question that in some settings, management opinion as to performance status of a DMU can be misdirected. This may result when management is focused on only one component of an operation, and fails to take full consideration of all aspects of performance. We argue, however, that in the problem setting considered herein, branch consultant knowledge must be treated as being more than opinion. Rather, it should be seen as expert knowledge, on par with the type of expertise modeled in any expert system. Evaluation of branch performance by internal consultants is a common practice in most major banks. Typically, micro-level work-studies are conducted within a sample of branches, to establish some form of standards. There is usually, however, no *transparent* definition of the mechanisms whereby the performance status of the branch is derived. This is generally due to the attempt by the consultant to merge any computed quantitative evaluation with qualitative factors that capture the environment or context within which the branch is compelled to conduct its business. This context can include the demographic makeup of the customer base, such as the financial profile of the average customer, age, ethnic makeup, and so on.

In most banks, there is seldom a single and definitive quantitative measure available as to the performance status of branches. Rather, the practice appears to be to classify branches into two or more groups on the basis of perceived *levels of productivity*. Arguably, incorporation of such information into a performance measurement model can serve to provide a more accurate representation of branch efficiency. As well, any model that builds on such information is more likely to succeed in being accepted internally.

3. *There is no clear reason why the central tendency focus of discriminant models should lend itself to aid in variable selection for frontier-based tools such as DEA.*

This position is difficult to dispute. It is the purpose of the experiments conducted herein, however, to provide evidence that, despite the obvious logic in this contention, these (central tendency) tools can, in certain situations, improve DEA performance in the sense of creating scores that reflect the expert's view.

The branch consultant's knowledge in the present setting is represented by a data set in which 200 bank branches are organized into two groups: 100 high performing and 100 low performing branches. A sample of this data is provided in the appendix. For each branch, data have been provided as discussed previously.

We do not attempt, in this particular instance, to reduce the number of variables by using a pre-screening process such as factor analysis or other statistical means. We are assuming that the branch consultants have selected the variables on which they wish to apply strategies, and that all of these variables have been deemed as important, and should, therefore, be retained. However, in many settings, such pre-screening would be essential. We assume here, as well, that all variables are flexible, and can be deemed as either inputs or outputs. This assumption is removed in the next section.

5.4.1 Linking Discriminant Techniques and the Additive DEA Model

In the additive DEA model of Charnes et al. (1985), the objective is to maximize the production of outputs for the minimum amount of inputs. This model has the advantage that the objective function is a summation of inputs and outputs ($\sum$ Outputs-$\sum$ Inputs). Recall that the formulation of the additive model is expressed by:

$$
\begin{aligned}
&\max \mu Y_0 - vX_0 \\
&\text{subject to:} \\
&\mu \mathrm{Y}_{\mathrm{k}} - vX_k \leq 0 \quad \forall_k, \\
&\mu \geq 1, \\
&v \geq 1.
\end{aligned}
$$

Thus, we can better understand the hypothesis stating that the selection of inputs and outputs can be based on the sign of the discriminant analysis coefficients. To see this, consider the logistic regression (LR) technique discussed earlier. Using the LR model of (5.1) the associated discriminating rule can be stated as follows:

If $\hat{P}_k \leq P_T$, then $\mathrm{DMU}_{\mathrm{k}} \in G_1$, else $DMU_k \in G_2$, where P_T is the LR threshold (usually 0.5).

The LR function can be restated as

$$P_k = \frac{1}{[1+\exp(-\sum_{q=1}^{Q} b_q z_{qk})]}$$

and LR rule can be divided into two inequalities

$$\hat{P}_k \le P_T,\ \text{DMU}_\text{k} \in G_1, \qquad (5.3a)$$

$$\hat{P}_k \ge P_T,\ \text{DMU}_\text{k} \in G_2 \qquad (5.3b)$$

Note that the formulae (5.3a) and (5.4b) resemble those of the linear goal programming discriminant model discussed earlier. In that case, the logistic regression threshold plays the role of the goal programming threshold. Let us define the function μ as the following linear combination:

$$u_k = \sum_{q=1}^{Q} b_q z_{qk}.$$

Then, we can restate discriminant equations (5.3a) and (5.3b)

$$\frac{1}{[1+\exp(-\mu_k)]} \le P_T,\ \text{DMU}_k \in G_1,$$

$$\frac{1}{[\ +\exp(-u_k)]} \ge P_T,\ \text{DMU}_k \in G_2$$

or equivalently as

$$u_k \le -\ln\left(\frac{1}{P_T}-1\right),\ \text{DMU}_k \in G_1, \qquad (5.4a)$$

$$u_k \ge -\ln\left(\frac{1}{P_T}-1\right),\ \text{DMU}_k \in G_2, \qquad (5.4b)$$

Notice that $-\ln(1/P_T-1)$ is the Cutoff Value 'T' for the LR model, when the function is linearized.

In contrast with the linear goal programming model, the LR model does not use a large value M to re-classify the observations on the wrong side of the hyperplane. Therefore, we can say that the final formulation (5.4a),(5.4b) is similar to the linear goal programming model with the following assumptions:

$$\text{T} = \text{-}\ln\left(\frac{1}{P_T}-1\right).$$

In solving LR model (5.1) (or GP model (5.2)), let Q_1 denote those variables $q=1,\ldots,Q$ for which $b_q<0$, and Q_2 those for which $b_q>0$. As well, use the notation

$$\mu_q = b_q,\ q\varepsilon Q_1,\ v_q = -\,b_q,\ q\varepsilon Q_2.$$

$$y_{qk} = z_{qk},\ q\varepsilon Q_1,\ x_{qk} = z_{qk},\ q\varepsilon Q_2.$$

Then, the expression for $u_k = \sum_{q=1}^{Q} b_q z_{qk}$ can be represented by

$$u_k = \sum_{q \varepsilon Q_1} \mu_q y_{qk} - \sum_{q \varepsilon Q_2} v_q x_{qk}$$

or

$$u_k = \mu Y_k - vX_k$$

in vector notation.

Comparing this to the format for the additive DEA model given above, one can immediately see the rationale for choosing as outputs Y_k, those variables in Z_k which are assigned positive coefficients, and as inputs X_k, those that receive negative coefficients.

In conclusion, with the additive DEA model it appears that the signs of the discriminant function (e.g., LR) coefficients can aid in determining the appropriate orientation of the variables: *a positive coefficient indicates an output, and a negative coefficient, an input.*

We do not concern ourselves here with significance of variables in the usual sense, as we assume all variables are to be retained.

5.4.2 Data transformation

A new data set has been included in this study to be used with goal programming and integer linear programming computations. This data set is a transformation of the original data set that has been provided by the bank. It is a projection onto an arbitrary positive interval (100,200), resulting in every observation being measured on the same scale. This projection avoids the ill-conditioned matrix phenomenon when using goal programming or integer goal programming. The projected data set has been used in the experiment to compare the results produced by these two discriminant techniques.

5.4.3 DEA measures

The DEA software used for the experiments is IDEAS V5.1. This software computes different efficiency scores depending on the model (additive, input, output, CRS, VRS). This study uses IOTA and DELTA measures computed. IOTA is a Ratio Measure, while DELTA is a Distance Measure.

Delta is a weighted aggregation of the differences between the observed and the projected points. If the observed point and the projected point are the same (efficient DMU), Delta will be zero. Delta is optimized for additive models.

5.5. THE SIMULATION EXPERIMENT

5.5.1 Methodology

We wish to test the claim that selecting the inputs and the outputs in a manner consistent with expert judgment, can actually improve the predictive capability of DEA models. The demonstration's aim will be to compare the DEA models thus improved with those not constructed with our method. The methodology consists of: (1) determining an average performance of the improved DEA models and, (2) comparing this average performance with the set of other DEA model's performances. The possible number of DEA models for a given type corresponds to the set of the possible combinations of variables; in our case, there are 728 DEA models (3^6 - 1). (It is noted that although in the current setting, only the $2^6 - 1 = 63$ full variable combinations are of interest, we display results for all combinations here).

The performance measure of quality that will be used to assess the expert DEA model, will be that model's ability to classify branches according to their DEA scores. A DEA model i will be considered as superior to model j, if model i properly classifies (consistent with management's judgment) more branches than is true of model j.

The average performance computation will be based on the use of a statistical methodology. Its principle is to calculate a performance that is the average of at least 10 similar experiments' performances. These experiments are similar as they all use the same initial data set. Each experiment builds a predictive model by using 90% of this initial data set and then testing this model on the 10% left. These subsets are created randomly, and the result of each predictive model indicates the performance of the corresponding experiment.

These experiments will also enable us to compare different techniques that help to define the inputs and outputs (i.e. LR, GP, IGP, MDA).

Therefore, the methodology used for this experiment is divided into two phases:

- The analysis phase, composed of five sequential stages, creates 10 predictive models using 10 analysis samples;
- The predictive phase tests these predictive models on 10 holdout samples.

5.5.2 Estimating the Predictive Model

The *Random Sampling Stage (1)* creates (randomly) 10 analysis and 10 holdout samples, out of the initial branch data set. These samples are used by the discriminant techniques through the remaining stages. The analysis samples have 180 observations each (90% of the branch data set), and the holdout samples have 20 observations each (10% of the branch data set). In each sample, half of its observations are classified as efficient, and the other half as inefficient to respect the initial proportions.

During the *Discriminant Techniques Stage (2),* the selected tools (LR, GP, IGP, MDA) are applied to the analysis samples. The signs of the coefficients for each discriminant function found, determine a combination number used to retrieve a DEA score in the Matching DEA Scores Stage. See Stage 4 below.

The *DEA Combinatorial Process Stage (3)* computes the DEA measure for every input/output combination (728 combinations) of each scenario. This process creates 10 sorted classification tables summarizing, for each combination, the best threshold and the number of properly classified observations. These classification tables are used in step (4) to find specific results indexed by a combination number obtained in step (2).

The *Matching DEA Scores Stage (4)* builds other tables summarizing DEA scores by technique. It also ranks the results according to a position within the DEA classification table.

The *Classification Summary Table Stage (5)* is the final stage before the interpretation of the analysis stage. It consolidates and summarizes the scores by averaging the results to obtain an average performance.

Finally, the *DEA Predictive Classification Stage (6)* (see next subsection) computes DEA scores on the holdout samples and classifies the

200 observations by using the thresholds found in the analysis process and computed on only 180 observations. During this stage, similar classification tables are created. Unlike other statistical experiments that would use only the holdout sample, here we are using the entire data set.

Table 5-2 summarizes the classification results for each discriminant technique. Each row indicates the average classification results. The first row, for example, displays the results computed when using logistic regression coefficient signs to determine inputs and outputs. The first column displays the average performance of the DEA models using these variable combinations to classify branches according to their DEA scores. The second column indicates the percentage of branches that are misclassified. The third column shows the average rank of this score within the DEA Classification Tables. The final column displays the number of properly classified DMUs in the holdout sample of 20 units. For example, the DEA Score of 136, in the first row, means that when using the logistic regression coefficients and the Delta measure (additive DEA score), 136 branches out of 180 are properly classified on average, meaning that 24.4% of the branches are misclassified. This result is, on average, in the 78.5th position(out of 728) within the DEA Classification Table. The last row shows the best results; in this case it indicates that determining inputs and outputs with goal programming is the best method for computing DEA scores in terms of classifying the branches in the best manner (the best results are underlined). It is noted that rescaling the original data gives the same results when using goal programming.

Table 5-2. Summarized DEA classification table (additive model/analysis stage)

Average results	Prop. Class (180)	% not class.	Ranking out of 728
Inputs/outputs (LR)	136	24.4	78.5
Inputs/outputs (GP)	145.5	19.2	50
GP (data rescaled)	145.5	19.2	50
Inputs/outputs (IGP)	128	28.9	149.5
IGP (data rescaled)	134.5	25.3	83
Inputs/outputs (MDA)	124	31.1	197
Best results	145.5	19.2	50

5.5.3 Testing the predictive model

The DEA Predictive Classification Stage is the second part of the methodology. The principle is to use the same combinations of inputs and outputs, and the same thresholds as those computed in the analysis part. This predictive process is divided into three steps.

- **Step 1:** The DEA Process computes the DEA scores for each possible combination of inputs and outputs (728 combinations). Note that we do not compute DEA scores on just the holdout samples, but on the entire data set composed of 200 branches. By using the thresholds and the coefficients computed in the previous part, we create 10 DEA classification tables.
- **Step 2:** The Matching DEA Scores Process is similar to the one applied during the analysis part. It consists of building one table per discriminant technique, indicating for each scenario the classification score that corresponds to the combination of inputs and outputs computed during the analysis phase.
- **Step 3:** The classification Summary Tables Process is the final stage that summarizes the measures for the 10 scenarios and the discriminant techniques into one table.

Table 5-3 shows the classification results obtained at the end of this process. Each row indicates the average classification results. The first row displays the results computed when using logistic regression coefficient signs to determine inputs and outputs. For example, the score of 160.5, in the second row, means that when using goal programming coefficients and the Delta measure (additive DEA score), 160.5 branches out of 200 are properly classified on average, which is similar to saying that 19.8% of the branches are misclassified. This result is, on average, in the 46.5th position (out of 728) within the DEA Classification Table. The final two columns present the outcomes from the holdout sample of 20 DMUs. For example, using GP to select variables for the DEA analysis, the resulting DEA model properly classifies approximately 16 out of the 20 hold out branches. Again, it would appear that goal programming is the best technique for selecting variables for computing DEA scores, in that it classifies the branches in the best manner (the best results are underlined). Specifically, goal programming appears to be a favorable vehicle for incorporating expert opinions into the DEA framework.

Table 5-3. Summarized DEA classification table (additive model/predictive stage)

	Out of 200	% Not	Out of 728	Out of 20	% Not
With LR	152	24.0	75	14.3	28.5
With GP	160.5	19.8	46.5	16.3	18.5
GP data rescaled	160.5	19.8	46.5	16.3	18.5
With IGP	138	31.0	206	14.2	29.0
With IGP rescaled	149.5	25.3	76	13.7	31.5
With MDA	149.5	25.3	76	13	35.0
Best results	160.5	19.8	46.5	16.3	18.5

5.6. VARIABLES WITH IMPOSED INPUT AND OUTPUT STATUS

The branch consultant knowledge used for these experiments is in the form of a classification of branches into two groups: the high and low performing branches. One can also take into account another type of information, namely predefined variable orientations. Indeed, in many cases, the branch consultants, when classifying branches, already know which variables are definitely inputs and which are outputs, versus those that can be considered as *either inputs or outputs*. It is important to consider this kind of information during the analysis.

In our case, the branch consultants were requested to specify which variables they would consider as inputs and which as outputs. They defined

FTETOT as being an input and RSP and LOANTOT as being outputs. They displayed no strong opinion about the remaining variables. We refer to these as flexible variables.

Table 5-4. Variable predefined orientations

Variable	Orientation	Description
FTETOT	Input	The sum of all full-time employees (sales/service positions
RSP	Output	The number of retirement savings plans sold
LOANTOT	Output	The total of all loans/mortgages
MOPCAO	Flexible	The total of accounts opened
MDPMTRF	Flexible	The number of deposits/transfers
MWDMUPD	Flexible	The number of withdrawals/updates

Therefore, this type of information can be incorporated into the models by matching the signs of the coefficients according to their predefined orientations. For instance, FTETOT is defined as an input, which means that its associated coefficient, within any discriminant model, should be negative. Similarly, an output variable indicates a positive coefficient.

There is no convenient mechanism for adding this kind of constraint to models such as logistic regression or multiple discriminant analysis, whereas this can be done with goal programming models. Fortunately, as discovered earlier, goal programming provides results that are approximately on par with logistic regression. Hence, there is no sacrifice in discriminant power, by resorting to goal programming as the tool of choice. Therefore, adding sign restriction constraints to the previous goal programming structure (5.2) results in the following formulation:

$$\begin{aligned} &\min \sum_{k=1}^{n} h_k \alpha_k \\ &\text{subject to:} \\ &\sum_{q=1}^{Q} b_q z_{qk} - \alpha_k \le T, \quad k \in G_1, \\ &\sum_{q=1}^{Q} b_q z_{qk} + \alpha_k \ge T, \quad k \in G_2, \\ &b_q \ge 0, \quad q \varepsilon Q^+, \\ &b_q \le 0, \quad q \varepsilon Q^-, \end{aligned} \tag{5.5}$$

where Q^+ is that subset of factors q in $\{1,...Q\}$ which are to be designated as outputs, and Q^-, those designated as inputs, In the current setting

$$Q^+ = \{\text{RSP,LOANTOT}\} \qquad Q^- = \{FTETOT\}.$$

The principle is to carry out the same experiments as those of the previous sections, but with three main differences:

- We use only one discriminant technique, goal programming, to determine the orientation of the flexible variables.
- Additional constraints are imposed in the model to take into account the fact that RST and LOANTOT are outputs, and that FTETOT is an input.
- The number of possible input/output combinations is reduced with three variable orientations now known; there are now 27 combinations.

Table 5-5. Additive DEA classification table for scenario #1

Combination nos. Out of 728	Input/Output Combination	DEA Threshold	Nos. Of branches Prop. Class	Nos. Of Non Prop. Class Branches
135	1 223 3	239	151	29
115	1 221 3	7561	150	30
109	1 221 1	7711	149	31
130	1 223 2	2766	148	32
133	1 223 3	4095	148	32
112	1 221 2	5011	145	35
127	1 223 1	4433	144	36
114	1 221 2	2841	135	45
117	1 221 3	3863	134	46
111	1 221 1	3780	133	47
132	1 223 2	592	131	49
129	1 223 1	577	129	51
124	1 222 3	267	126	54
118	1 222 1	2272	124	56
121	1 222 2	2392	124	56
123	1 222 2	3070	117	63
120	1 222 1	2421	116	64
126	1 222 3	3168	116	64
113	1 221 2	950	112	68
116	1 221 3	1854	109	71
122	1 222 2	4631	109	71
131	1 223 2	2459	108	72
125	1 222 3	5109	107	73
119	1 222 1	5137	106	74
110	1 221 1	2796	105	75
134	1 223 3	2300	103	77
128	1 223 1	1756	99	81

According to the branch consultant strategy, some input and output combinations will not be possible even if they are included in the 27 remaining cases. Indeed, if one looks at Table 5-5, it can be seen that the first row indicates the combination #135 that corresponds to the following variable combinations: one input (FTETOT), two outputs (RSP and

LOANTOT), with the other variables are not considered for the analysis. Recall that the strategy is to keep all variables within the analysis scope to be able later on to reduce inputs or increase outputs of the inefficient branches to bring them to the efficient frontier. Therefore, the ranking will be based on the remaining combinations.

Table 5-5 displays for one of the ten scenarios (i.e. scenario #1), the input and output combinations and the classification results for each of the 27 possible combinations. The values in this table are sorted, in descending order, according to the number of properly classified branches. The best DEA model (in terms of classification capability) is in first position, and the worst is in last position. Similar tables have been obtained for the other scenarios, but have not been displayed here. Each row of this table specifies:

- Its input and output combination number.
- Its input and output combination description. A value of 1 defines an input, a value of 2 an output, and a value of 3 that the variable is not included in the analysis. For instance, combination #109 indicates designations.

Variable	Combination Id	Orientation	Comment
FTETOT	1	Input	Predefined
RSP	2	Output	Predefined
LOANTOT	2	Output	Predefined
MOPCAO	1	Input	Flexible
MDPMTRF	1	Input	Flexible
MWDMUPD	1	Input	Flexible

- The computed DEA threshold to discriminate the branches.
- The number of properly classified branches (out of 180), using this input and output combination, and this DEA threshold;
- The number of non-properly classified branches.

It is noted that combination #109, the goal programming optimal combination, ranks in 3rd place out of the 27 combinations considered. In addition, if we consider that, in accordance with bank consultants' strategy, we want to keep every variable in the analysis, combination #109 is, in fact, in second position. Indeed, combination #135 should be excluded from the analysis set since the remaining flexible variables are not included in the analysis (they are neither inputs nor outputs).

The goal programming model presented above has been applied to each of the ten scenarios. The objective of this process is to determine 10 input and output combinations that will be used to find the corresponding DEA classification tables computed previously.

Table 5-6 displays summaries of the experiments of the additive oriented DEA model. The ten-fold cross validation methodology has been used to compute an average performance for the restricted models. The table has 10 rows, one for each subset of 180 branches. Each subset is used in the goal programming restricted model to find the input/output combination. In that case, combination #109 is chosen for every scenario (it could be different for some of them). Each row, then, indicates the number of branches properly classified, and the percentage of branches not properly classified, when using the DEA scores computed by the model, (additive or input), that corresponds to the inputs and outputs defined by combination #109. The last column displays the ranking of each scenario, within the 27 sorted possible combinations (Table 5-5). Notice that the ranking does not exclude the scenarios with variables excluded (such as the scenario #135).

Table 5-6. Averaged additive DEA classification results for the 10 scenarios

Class results	DMUs out of 180	% not prop.classified	Ranking out of 27
109	149	17.2	3
109	147	18.3	3
109	150	16.7	2
109	147	18.3	3
109	141	21.7	4
109	147	18.3	3
109	146	18.9	4
109	148	17.8	3
109	151	16.1	2
109	145	19.4	4
Average	147.1	18.3	3.1

In conclusion, restrictions imposed on the goal programming model multipliers, to express variable orientations predefined by bank consultants, provides better results, on average, than is true of the unrestricted DEA version. Recall that the average performance of the additive experiment gave 145.5 branches properly classified while the average restricted result is 147.1 branches properly classified.

5.7. THE INPUT-ORIENTED MODEL

We now examine the incorporation of classification data into the input-oriented radial model of Charnes et al. (1978). Recalling that the linearized form of this model is given by

$$\max\ \mu^T Y_o$$
subject to:
$$v^T X_o = 1$$
$$\mu^T Y_j - v^T X_j \le 0, \text{ all } j$$
$$\mu_r, v_i \ge 0, \text{ all } r, i,$$

It is noted that one can equally write the objective function in the form $\max\ \mu^T Y_o - v^T X_o$, since $v^T X_o$ is a constant by virtue of the first constraint in the above model.

One can immediately see the connection with discriminant analysis here, wherein we designate what will be an output versus an input according to the signs of the coefficients of those variables in a discriminant model.

A simulation experiment identical to that described above was applied to the radial (input-oriented) model. Table 5-7 is analogous to the earlier Table 5-2.

Table 5-7. Summarized DEA Classification Table (Input/Analysis)

Discriminant Technique used	# of properly classified DMUs out of 180 using DEA scores	% of not properly classified DMUs	Position # out of 728 sorted DEA combinations
Inputs and Outputs selected with LR coefficient signs	162	10.0%	35.5
Inputs and Outputs selected with GP coefficient signs	156	13.3%	66.5
Inputs and Outputs selected with GP coefficient signs (Data Rescaled)	156	13.3%	66.5
Inputs and Outputs selected with ILP coefficient signs	146	18.9%	130.5
Inputs and Outputs selected with ILP coefficient signs (Data Rescaled)	159	11.7%	47.5
Inputs and Outputs selected with MDA coefficient signs	144.5	19.7%	138
Best Results	162	10.0%	35.5

In this case the LR model performs slightly better than the GP model as per Table 5-2

The DEA Predictive Classification Stage is the second part of the methodology. The principle is to use the same combinations of inputs and outputs, and the same thresholds as those computed in the analysis part. This predictive process is divided into three steps.

Table 5-8. Summarized DEA Classification Table (Input/Predictive)

Discriminant Technique used	# of properly classified DMUs out of 180 using DEA scores	% of not properly classified DMUs	Position # out of 728 sorted DEA combinations
Inputs and Outputs selected with LR coefficient signs	182.5	8.8%	26.5
Inputs and Outputs selected with GP coefficient signs	171	14/5%	62.5
Inputs and Outputs selected with GP coefficient signs (Data Rescaled)	171	14.5%	62.5
Inputs and Outputs selected with ILP coefficient signs	163	18.5%	120
Inputs and Outputs selected with ILP coefficient signs (Data Rescaled)	180	10.0%	32
Inputs and Outputs selected with MDA coefficient signs	153.5	23.3%	179
Best Results	182.5	8.8%	26.5

Table 5-8 shows the classification results obtained at the end of this process. Each row indicates the average classification results. The first row displays the results computed when using logistic regression coefficient signs to determine inputs and outputs. For example, the score of 182.5, in the second row, means that when using logistic regression coefficients and the Theta measure (input-oriented DEA score), 182.5 branches out of 200 are properly classified on average, which is similar to saying that 8.8% of the branches are misclassified. This result is, on average, in the 26.5th position (out of 728) within the DEA Classification Table. Here, it would appear that logistic regression is the best technique for selecting variables for computing DEA scores, in that it classifies the branches in the best manner (the best results are underlined). It should be pointed out, however, that goal programming appears to be nearly as favorable a vehicle for incorporating expert opinions into the DEA framework.

Table 5-9 displays the best classification scores of the additive and input-oriented DEA model experiments, at the end of the analysis stages. It is useful to note that the input-oriented model gives better classification results than those of the additive model. It is clear, however, that both models perform very well when using our theory to select variables. Indeed, if we look at the additive DEA model, the best classification score ranks at the 50th position within the 728 combinations, versus position 35 for the input-oriented model.

Table 5-9. Comparison of Analysis Stage Results

Comparison of the best DEA Classification Scores for the Additive and Input Oriented Models at the end of the Analysis stages			
	# of properly classified branches out of 180	% of non properly classified branches	Position #within 728 possible combinations
Additive DEA Model	145.5	19.2%	50
Input DEA Model	162	10%	35.5

From this comparative table, we can say that for either model, when utilizing the variables from the best discriminant tool, the results outperform those corresponding to most of the random combinations of inputs and outputs.

Table 5-10. Comparison of Predictive Stage Results

	# of properly classified branches out of 200	% of non properly classified branches	Position #within 728 possible combinations
Additive DEA Model	160.5	19.8%	46.5
Input DEA Model	182.5	8.8%	26.5

Table 5-10 displays the best classification scores of the additive and input-oriented DEA model experiments, at the end of the predictive stage. Here again, we can see that the classification results are impressive.

Performance of Holdout Sample

The models computed during the predictive stage have been applied to the entire set of 200 branches, instead of simply on the holdout sample of 20 branches. Due to the relative efficiency nature of DEA, the desire was to compare the improved DEA model with the basic one. However, to demonstrate the effectiveness of the idea, the results for the 20 branches in the 10 experiments were extracted and are displayed in Table 5-11.

Table 5-11. Classification of Holdout Samples Using Input DEA Models and Goal Programming Coefficient Signs to Select Inputs and Outputs

Holdout Samples	Threshold	# of properly classified DMUs Out Of 20 Using DEA Scores	% of Not Properly Classified DMUs
1	0.59213	16	20.0%
2	0.58852	16	20.0%
3	0.5661	17	15.0%
4	0.56224	16	20.0%
5	0.6749	17	15.0%
6	0.60775	17	15.0%
7	0.59785	15	25.0%
8	0.59826	18	10.0%
9	0.60775	19	5.0%
10	0.4578	16	20.0%
Average		16.7	16.5%

The results show that on average 83.5% of the holdout DMUs are properly classified in the input-oriented model.

Imposed Input and Output Status

As in the case of the additive model, the radial input-oriented model was examined when a subset of the available variables is already preclassified as inputs or outputs. Following the same analysis as conducted previously, Table 5-12 displays the results.

The format of this table is the same as described earlier. In conclusion, restrictions imposed on the goal programming model multipliers, to express variable orientations predefined by bank consultants, provides better results, on average, than is true of the unrestricted DEA version. Recall that the average performance of the input-oriented experiment gave 156 branches properly classified while the average restricted result is 164.5 branches properly classified.

5.8. GP CONSTRAINT – ENHANCED DEA MODEL

The model structures discussed above, build expert opinion into the *first* stage of the analysis, where classification models are applied to decide variable designation (inputs and output). In the second stage, a standard DEA model is used to derive performance scores for the DMUs. It can be argued that the performance measures can be enhanced by *re-introducing* the expert's classification information directly into the *second* stage DEA structure itself. Specifically, we permit the expert to intervene in this stage, by imposing constraints on the DEA model that capture his/her decisions. The hypothesis is that by integrating this additional knowledge into the model, the results will be more consistent with expert heuristics.

Table 5-12. Input DEA Classification Table for Scenario #1

Combination # out of 728	Input & Output Combination	DEA Threshold	# of properly classified branches out of 180 using DEA scores	# of non properly classified branches
135	12233	0.20576	179	1
109	12213	0.28851	165	15
130	12211	0.32661	154	26
132	12221	0.40456	151	29
117	12232	0.19817	150	30
123	12212	0.23472	148	32
124	12231	0.54461	148	32
126	12212	0.18983	147	33
133	12213	0.16233	147	33
127	12211	0.43612	146	34
112	12232	0.33523	145	35
111	12231	0.23081	144	36
121	12223	0.57055	144	36
134	12221	0.14155	144	36
114	12222	0.44277	143	37
125	12222	0.15762	143	37
122	12221	0.17537	142	38
131	12223	0.20618	141	39
115	12212	0.06281	138	42
129	12213	0.23995	138	42
118	12222	0.58899	136	44
120	12232	0.37177	134	46
113	12223	0.63637	132	48
116	12221	0.62134	132	48
119	12211	0.30946	128	52
128	12233	0.30817	126	54
110	12231	0.71518	116	64

The approach will be to compare the results arising from this enhanced model, on which additional constraints have been imposed, with those from a comparable non-restricted model. Again, let G1 be the set of the high performing branches of the analysis sample (90 branches) and G2 be the set of 90 low performing branches. The DEA model is applied on the entire data set of 200 branches, but we impose classification restrictions on a subset of the data set (180 branches).

Reconsider problem (5.6), but where we desire to impose additional goal programming constraints to discriminate between the branches of the two groups G_1 and G_2, according to their respective DEA scores. Here, the DEA measure is defined by the ratio: $u^T Y_o / v^T X_o$. Consider then, the following integer goal programming model.

$$\begin{aligned}
&\min \sum \gamma_i \\
&\text{subject to:} \\
&\frac{u^T Y}{v^T X} + M\gamma_i \geq T + \varepsilon, \quad G1 = \{1..90\} \\
&\frac{u^T Y}{v^T X} - M_{\gamma_i} \leq T - \varepsilon, \quad G2 = \{91..180\} \\
&\frac{u^T Y}{v^T X} \leq 1 \\
&u^T \geq \vec{0} \quad v^T \geq \vec{0} \\
&\gamma_i \in \{0,1\} \\
&T \text{ unrestricted}
\end{aligned} \tag{5.7}$$

Let $\hat{M}, \hat{\gamma}_i$, and $\hat{T}$ be the optimal values derived in this model.

This model attempts, through the first two constraint sets (on G_1 and G_2) to properly classify the members of G_1 and G_2. The variables γ_i record the number of misclassified branches.

This formulation is clearly nonlinear, and in real situations, with large data sets, deriving a solution can be computationally challenging. It can be important to avoid adding such nonlinear restrictions into the input oriented model. The purpose of the experiment herein is to compare three approaches:

(1) The non-restricted input oriented model, expressed in its linear version;
(2) The input oriented model, with nonlinear goal programming constraints added. The results computed by this enhanced DEA model takes into account expert opinion, expressed in a ratio form;
(3) The input oriented model, with linear goal programming constraints added. Indeed, with this experiment, we wish to demonstrate that this can be a good approximation of the nonlinear version (i.e. ratio constraints). The performance of this DEA model will be compared with the two other cases, to determine if it can be used as a replacement for the nonlinear version, and if it is providing better results than a non-restricted model.

5.8.1 Imposing Nonlinear Goal Programming Constraints in an Input Oriented DEA Model

The nonlinear problem presented above was solved, and the optimal values derived were inserted into the goal constraints, thereby transforming them into a linear form as follows (this holds, since the optimal values are now scalars):

$$\mu^T Y \geq (\hat{T}+\varepsilon-\hat{M}\hat{\gamma}_i)v^T X, \quad i \in G_1 = \{1..90\}$$
$$\mu^T Y \leq (\hat{T}-\varepsilon+\hat{M}\hat{\gamma}_i)v^T X, \quad i \in G_2 = \{91..180\}$$

These constraints are then added to the linear form of the input-oriented model:

$$\begin{aligned}
&\max \mu^T Y_0 \\
&\text{subject to:} \\
&v^T X_0 = 1 \\
&\mu^T Y - v^T X \leq 0 \\
&\mu^T Y \geq (\hat{T}+\varepsilon-\hat{M}\hat{\gamma}_i)v^T X \quad i \in G_1 = \{1..90\} \\
&\mu^T Y \leq (\hat{T}-\varepsilon+\hat{M}\hat{\gamma}_i)v^T X \quad i \in G_2 = \{91..180\} \\
&\mu^T \geq \vec{0} \quad v^T \geq \vec{0}
\end{aligned}$$

Table 5-13. Summarized Results for the Input Restricted Experiment

	Unrestricted (0)	A1 (1)	A2 (2)	A3 (3)	A4 (4)	A5 (5)	A6 (6)	A7 (7)	A8 (8)	A9 (9)	A10 (10)	Average A1-A10
Threshold	.44	.30	.30	.25	.30	.24	.22	.30	.33	.30	.29	.288
# of Properly Classified Branches	171	191	190	189	190	188	192	191	189	191	188	189.9
# of non Properly Classified Branches	29	9	10	11	10	12	8	9	11	9	12	10.1

This three-stage process, applied to the bank data set, is similar to the previous experiment, and Table 5-13 summarizes the results of the ten restricted DEA models compared to the unrestricted DEA model. The last column demonstrates that, on average, the nonlinear restricted version of the

DEA model is properly classifying 189.9 branches out of 200 branches. If we compare this average performance with the unrestricted version, we recognize that it is doing better, both in terms of classification, and in terms of overall benchmarking.

5.8.2 Imposing Linear Goal Programming Constraints in an Input Oriented DEA Model

This case assumes that the following linear goal programming model can be used in place of the nonlinear formulation. Notice that the third set of constraints indicates that every DEA score must not exceed 1. In fact, even if we used the net profit oriented form as an approximation to the ratio form we still need to keep the original requirements of the input oriented model. Hence, we have imposed the DEA constraints into the goal programming model, to respect the nature of the DEA scores:

$$\begin{aligned} &\min \ \sum \gamma_i \\ &\text{subject to:} \\ &\mu^T Y - v^T X + M\gamma_i \geq T + \varepsilon \quad G1 = \{1..90\} \\ &\mu^T Y - v^T X - M\gamma_i \leq T - \varepsilon \quad G2 = \{91..180\} \\ &\mu^T Y \leq v^T X \\ &\mu^T \geq \vec{0},\ v^T \geq \vec{0},\ \gamma_i \in \{0,1\} \\ &T \text{ unrestricted} \end{aligned} \tag{5.9}$$

Let $\hat{M}$, $\hat{\gamma}_i$ and $\hat{T}$ be the optimal values derived in this model, and consider the constrained DEA model

$$\begin{aligned} &\max \mu^T Y_o \\ &\text{subject to:} \\ &v^T X_0 = 1 \\ &\mu^T Y - v^T X \leq 0 \\ &\mu^T Y - v^T X \geq \hat{T} + \varepsilon - \hat{M}\hat{\gamma}_i \quad i \in G_1 = \{1..90\} \\ &\mu^T Y - v^T X \leq \hat{T} - \varepsilon + \hat{M}\hat{\gamma}_i \quad i \in G_2 = \{91..180\} \\ &\mu^T \geq \vec{0} \ \ v^T \geq \vec{0} \end{aligned} \tag{5.10}$$

Again, this process is similar to the one used with the additive formulation. The final stage computes the DEA model, and classifies the scores for each of the ten versions. Table 5-14 summarizes the results for the unrestricted version, and the ten restricted models. The last column is the average of the ten restricted results, and gives an estimate of the performance we can have with this approximation. We can see that 188.8 out of 200 branches are properly classified, while only 171 branches are properly classified with the unrestricted version. This average performance is very close to the nonlinear restricted performance (i.e. 189.9). Therefore, we can conclude that instead of using a nonlinear form for restricting an input oriented model, one can use its linear approximation, and get close results.

Table 5-14. Summarized Results for the Input Restricted Experiment (Linear Constraints)

	Unres. (0)	A1 (1)	A2 (2)	A3 (3)	A4 (4)	A5 (5)	A6 (6)	A7 (7)	A8 (8)	A9 (9)	A10 (10)	Avg. A1-A10
Threshold	0.44	0.34	0.31	0.26	0.30	0.25	0.22	0.31	0.30	0.31	0.29	0.2
# Prop Classified Branches	171	189	188	189	190	188	191	187	189	187	190	188.8
# non Prop Classified Branches	29	11	12	11	10	12	9	13	11	13	10	11.2

5.9. SUMMARY

This chapter has examined the embedding of expert knowledge within the DEA model structure. The principal form that such information will take is a classification of a subset of DMUs into two or more groups. We examine only the case of two groups or classes here. It has been demonstrated that considering such information can result in DEA scores that are more in line with management's view of performance.

Modifications of this idea can be executed, such as the imposition of sign restrictions on factor multipliers, arising from locking in such factors as either inputs/outputs. Finally, the DEA analysis can be further enhanced by requiring that those DMUs classified by management, obey that classification to the extent possible within the wider DEA analysis.

REFERENCES

1. A. Charnes, W.W. Cooper, B. Golany, L.M. Seiford, and J. Stutz, (1985), Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions, *Journal of Econometrics* 30, 91-107.
2. A. Charnes, W.W. Cooper, and E. Rhodes (1978), Measuring the

efficiency of decision making units, *European Journal of Operational Research* 2, 429-444.
3. W.D. Cook, A. Kazakov, and Y. Roll, (1990), A DEA model for measuring the relative efficiency of highway maintenance patrols. *INFOR* 28, 113-124.
4. W.D. Cook and J. Zhu (2003), Output deterioration with input reduction in DEA, *IIE Transactions* 35, 309-320.
5. W.D. Cook and K. Bala (2003), Performance measurement with classification information: an enhanced additive DEA model. *OMEGA* 31, 439-450.

This chapter is based upon Bala, K. and Cook, W.D., 2003, Performance Measurement with Classification Information: An Enhanced Additive DEA Model, *OMEGA*, 31:439-450, with permission from Elsevier Science.

Chapter 6

MULTICOMPONENT EFFICIENCY MEASUREMENT IN BANKING

6.1. INTRODUCTION

Banks have evolved over time from their traditional role as reactive monetary intermediaries, and *service* providers, toward a more general and proactive function as universal financial agents with a distinct *sales* culture. This new status has resulted in the introduction of a broad range of financial products to the market place. Under the Canadian Bank Act of 1991, it became legal for an institution to engage in a broad range of financial activities. Technology has contributed as well to the changes that banks are undergoing; a range of convenient customer access points has emerged such as ATMs (Automatic Teller Machines), debit cards, telephone- and PC banking, to name a few.

Banks generate profits from two main sources -- (1) interest income, which captures the spread realized on loans and traditional activities, and (2) non-interest income from fees and financial services activities. While historically interest income was the principal source of profits for the bank, the importance of non-interest income has grown significantly over time. It is interesting to note that the profitability ratio, that is the profit as a percentage of assets, has increased dramatically since 1991. Specifically, for the period 1980—1990, the ratio ranged from 0.24% to 0.79%, with an average of 0.43%; the corresponding figures for the period 1991-1995 are 0.59% to 1.90% with an average of 1.20%. This dramatic change has been due in part to the revised regulations in the Bank Act, and partially to improved access to financial services, coupled with a more active sales orientation.

Performance measurement, using tools such as Data Envelopment Analysis (DEA), as proposed by Charnes et al. (1978), has tended to concentrate on achieving a *single measure* for each member of a set of decision making units (DMUs). In most applications, a *single* measure of production or profit efficiency provided by the DEA methodology has been an adequate and useful means of comparing units and identifying best performance. This has been particularly true in the case of banks, where the primary candidates for DMUs are branches, and in their traditional setting, product and prices have tended to be undifferentiated. Numerous studies of bank-branch efficiency using DEA have been conducted over the past 15 years – see, for instance, Charnes et al. (1990), Oral and Yolalan (1990), Schaffnit et al. (1997), Sherman and Gold (1985), and Sherman and Ladino (1995).

There is now a desire to create value-added customer segments by identifying their specific needs. The new challenge is to optimize resource allocation, with most of the industry now allocating 60-80% of its human capital to customers and markets that represent less than 20% of its customer base. There is a growing need to view performance in a more dis-aggregated sense, paying specific attention to different components of the operation. These components include different classes of products or sales activities, such as mutual funds and mortgages, and different elements of service. By measuring a branch's performance on each of a set of such components, particular areas of strength and weakness can be identified and addressed, where necessary.

In this chapter we present models for deriving *aggregate* measures of bank-branch performance, with accompanying *component* measures that make up that aggregate value. The technical difficulty surrounding the development of an appropriate model has to do with the presence of *shared resources* on the input side, and mechanisms for allocating such resources to the individual components.

The idea of measuring efficiency relative to certain subprocesses or components of a DMU is not new. Färe and Grosskopf (1996), for example, look at a multistage process wherein intermediate products or outputs at one stage, can be both final products and inputs to later stages of production. Those authors are not explicitly interested in obtaining measures of efficiency at each stage, but rather are concerned with overall efficiency measurement, whereby the network structure of the intermediate activity explicitly enters into the model description. Hence, they are able to provide a better representation of the technology than would a 'black box' input and final output model. Another example is due to Färe and Primont (1984) and involves the evaluation of efficiency of a set of multiplant firms as DMUs, while at the same time measuring the efficiency of plants within firms.

These applications of multicomponent efficiency measurement do not involve shared resources as does the situation examined herein. The work of Beasley (1995) on separating teaching and research, most closely compares to the present application, although we show herein that our treatment of shared resources leads to a linear rather than a nonlinear model. Section 6.2 modifies the conventional radial projection DEA model for bank-branch performance by providing a methodology for splitting shared inputs among the identified components. For development purposes, we concentrate on two specific components, namely service-specific and product-specific sales activities. The model structure used is based on the original CRS model of Charnes et al. (1978). An application is examined in Section 6.3. In Section 6.4 we present an additive form of the multi-component model. Discussion and conclusions follow in Section 6.5.

6.2. A MULTICOMPONENT PERFORMANCE MEASUREMENT MODEL

With the increased emphasis on sales and the differentiation of products and customer segments, there is a need to provide a performance measurement tool with component-based information as part of the aggregate efficiency score.

6.2.1 Multiple Functions and Shared Resources

While one may wish to measure the performance of *several* components of the DMU, we will, for purposes of development in this chapter, assume that transactions can be separated into exactly *two* distinct classes: service and sales. It should be emphasized that this split is not always transparent; the opening of a mortgage loan would generally be classified as a "sales" transaction, although there are "service" activities that must be performed from time to time pertaining to that loan, such as loan renewal. Thus, a particular transaction may contain both sales and service components. Care should, therefore, be exercised in clearly delineating those activities that belong to each function. Furthermore, one would generally need to separate those sales activities that are volume related (and pertain to specific products), from those that involve the "selling" part of the sales activities. The latter would include reviewing customer portfolios, answering customer requests on various products, and so on. The former would involve the transaction tasks performed after the customer has chosen a particular product. In summary, the selling aspect of sales does not relate to specific sales products while the transaction part of sales is product-specific. In this section we consider only those sales activities that are product or volume specific. We take up the non-volume related activities in a later section.

For notational purposes, let $\left(Y_j^1, Y_j^2\right)$ denote the sets of service and sales transactions, respectively, i.e. the two sets of *outputs* are

$$Y_j^1 = \left(y_{j1}^1 ..., y_{jJ_1}^1\right) \text{ and } Y_j^2 = \left(y_{j1}^2 ..., y_{jJ_2}^2\right).$$

On the *input* side, this split is more complex. Some resources can be designated as *dedicated* service inputs, some as dedicated to sales, and still others are *shared* by the two functions. If, for example, branch staff are classified as Sales, Service, and Support, we can, for illustrative purposes, assume that Support staff are shared by the two functions while the other two classes are dedicated. In some branches this distinction may be less clear than in others. Technology resources may as well be classified as shared.

A schematic of the *production process* for a particular DMU is given in Figure 6-1.

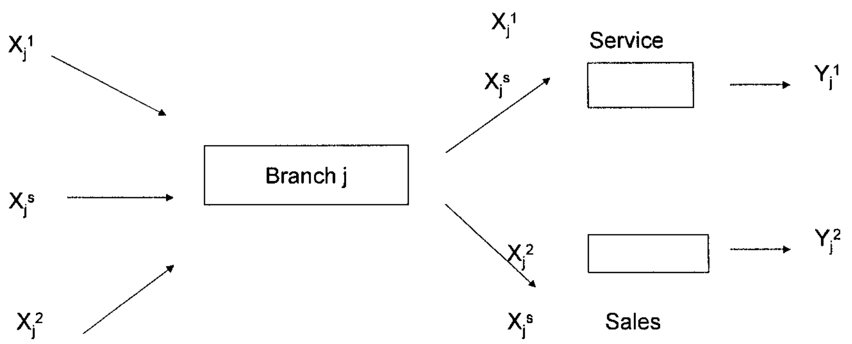

Figure 6-1. Production Process for a DMU*j with Shared Resources*

Here, X_j^1, X_j^2 and X_j^s denote I_1, I_2 and I_s-dimensional vectors of service dedicated-, sales dedicated-, and shared inputs, respectively. Some portion $\alpha_i \left(0 \leq \alpha_i \leq 1\right)$ of the shared resource x_{ij}^s is allocated to the service function of DMU j, with the remainder $\left(1-\alpha_i\right)$ being allocated to sales. In the model to be developed herein, α_i is a decision variable to be set by the DMU. At least two difficulties arise in attempting to capture a measure of performance of the DMU on both service and sales functions within some *overall* efficiency measure. First, if one attempts to derive an overall measure of performance that somehow incorporates sales and service components, the importance of the components of X^s relative to one another, and relative to the dedicated resources X^1 and X^2 (as reflected in the v-vectors v^1, v^2 and v^s), may be different when considering the impact of X^s on Y^1 as compared to its impact on Y^2. For example, consider the simple case of one staff type for each dedicated class (X^1 = no. service staff, X^2 = no. sales staff), and two resources, support staff and available technology, as shared inputs. One may argue that in evaluating service efficiency, technology is more important than support staff. As an example, a constraint such as $v_2^s \geq 2v_1^s$ might be imposed. On the other hand, if technology such as ATMs play a minor role in sales, then a constraint such as $v_2^s \leq 0.3v_1^s$ may be an accurate reflection of the importance of the two shared resources relative to one another. Clearly, these constraints are infeasible if imposed simultaneously. Moreover, even if this issue could be resolved, there would be no clear way of separating the resulting aggregate measure into separate sales and service indicators.

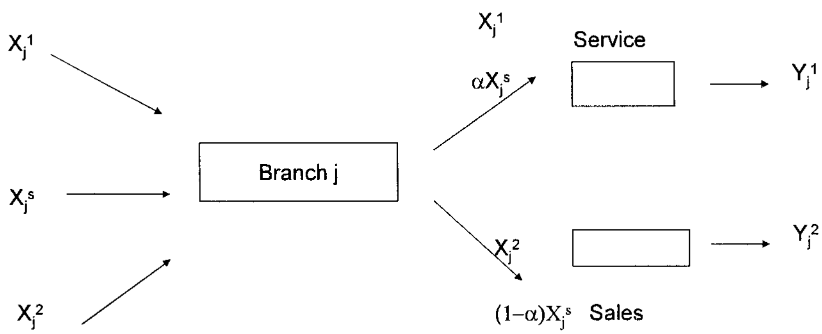

Figure 6-2. Splitting Shared Resources

A second difficulty arises if instead of developing an aggregate measure, one attempts to derive separate measures of performance relative to sales and service, with the intention of combining these separate measures into an aggregate score after the fact. The problem here is that the shared resources X^s would need to be *apportioned* to these two functions in some manner consistent with their usage in creating the outputs of the functions. With any shared resources, however, branches do not generally maintain a record of the usage split at the function level. Consequently, a mechanism is needed to split shared resources across functions in some equitable manner. To motivate the development, reconsider Figure 6-1, but with the shared resources X_j^s allocated to the two functions according to *proportionality variables*, α_i as depicted in Figure 6-2. The issue of how α_i should be derived is discussed below. Let $\alpha = (\alpha_1, \alpha_2, ... \alpha_{I_s})$ denote the column vector of proportionality variables, and let αX_j^s denote the column vector

$\left(\alpha_1 x_{1j}^s, \alpha_2 x_{2j}^s, \ldots \alpha_I x_{jI}^s\right)^T$. Further, we let $(1-\alpha)X_j^s$ denote the column vector $\left(\left(1-\alpha_1\right)x_{1j}^s, \left(1-\alpha_2\right)x_{2j}^s, \ldots, \left(1-\alpha_{I_s}\right)x_{jI_s}^s\right)^T$.

6.2.2 The Aggregate Performance Measure

From Figure 6-2 one can argue that since the total bundles of outputs Y_j^1 and Y_j^2 are produced from the inputs X_j^1, X_j^2 and X_j^s, a measure of aggregate performance e_j^a can be represented by:

$$e_j^a = \frac{u^1 Y_j^1 + u^2 Y_j^2}{v^1 X_j^1 + v^{s_1}\left(\alpha X_j^s\right) + v^{s_2}\left(\left(1-\alpha\right)X_j^s\right) + v^2 X_j^2} \tag{6.1}$$

For this representation, the vectors of multipliers u^ℓ and v^t would be determined in a DEA manner to be discussed below. The rationale for allowing for the possibility of different vectors v^{s_1} and v^{s_2} for the shared service and sales resources, respectively, is that the relative importance of the components of X^s in generating Y^1 may be different than their importance in generating Y^2. This was discussed earlier. In this manner, we avoid the possibility of infeasibilities created by possibly conflicting restrictions on the multipliers v^s. There is yet another rationale for permitting v^{s_1} and v^{s_2} to be different multiplier vectors. It can be argued that normally in a DEA analysis there is no clear connection between subsets of outputs and subsets of inputs. In this event, it is certainly the case that v^{s_1} and v^{s_2} should be the same vectors since they pertain to the same inputs (for example, support staff). When a direct link can be made between such subsets of input and output bundles, however, one might then attempt to impose some form of linking constraints as discussed in earlier literature. We do this in the model discussed below. Such constraints may only be feasible if v^{s_1} and v^{s_2} are, in fact, permitted to be different vectors.

6.2.3 Function-Specific Performance Measures

From e_j^a, performance measures for DMU j that capture service and sales efficiency would appear to be appropriately represented by e_j^1 and e_j^2, respectively, as defined by:

$$e_j^1 = \frac{u^1 Y_j^1}{v^1 X_j^1 + v^{s_1}\left(\alpha X_j^s\right)} \tag{6.2}$$

and

$$e_j^2 = \frac{u^2 Y_j^2}{v^{s_2}\left((1-\alpha)X_j^s\right) + v^2 X_j^2} \tag{6.3}$$

Property 6.1 *The aggregate performance measure* e_j^a *is a convex combination of the service and sales measures.*

Specifically $e_j^a = \beta_j e_j^1 + (1-\beta_j) e_j^2$, where β is the portion of all inputs utilized in e_j^1 (applied to the service component), i.e.

$$\beta_j = \frac{\left[v^1 X_j^1 + v^{s_1}\left(\alpha X_j^s\right)\right]}{\left[v^1 X_j^1 + v^{s_1}\left(\alpha X_j^s\right) + v^{s_2}((1-\alpha)X_j^s) + v^2 X_j^2\right]}.$$

The aggregate measure is, therefore, a weighted average of the performance across the various functions of the organization, as one would intuitively expect. From this property it is seen that a DMU will be deemed efficient, if and only if it is efficient in both service and sales components. Again we point to the importance of *separate* vectors v^{s_1}, v^{s_2} being permitted in the aggregate measure (6.1). If v^{s_1} and v^{s_2} are forced to be the same in (6.1), yet are permitted to be different in (6.2) and (6.3), then no connection between the aggregate and function-specific measures, as per Property 1, can be made.

6.2.4 Derivation of e_j^a, e_j^1, e_j^2

The defined measures are based upon proportionality variables α which will be treated as DMU-specific variables. Thus, it will be at the discretion of each DMU j to allocate X_j^s across the two functions. Furthermore, the model will make the necessary provisions to ensure that all three measures are appropriately scaled, specifically they will not exceed unity.

Consider the following mathematical programming model:

$$\begin{aligned}
&\max e_o^a \\
&\text{subject to:} \\
&e_j^a \le 1 && \forall j \\
&e_j^1 \le 1, && \forall j \\
&e_j^2 \le 1, && \forall j \qquad (6.4) \\
&0 \le \alpha_i \le 1, && \forall i \\
&(\mu^1, \mu^2) \in \Omega_1 \\
&\left(v^1, v^2, v^{s_2}, v^2\right) \in \Omega_2 \\
&u_r^1, u_r^2, v_i^1, v_i^2, v_i^{s_1}, v_i^{s_2} \ge \delta, && \forall i, j
\end{aligned}$$

In this formulation, the objective is to maximize the aggregate efficiency rating for each DMU "o", while ensuring that the function level ratings (for sales and service) do not exceed 1. We replace ε by δ here to denote the fact that an absolute lower bound δ may be in effect. The sets Ω_1 and Ω_2 are *assurance regions* (see Thompson et al. 1990) defined by any restrictions imposed on the multipliers. Similar work was done by Beasley and Wong (1990). The set Ω_1 may, for example, contain ratio constraints on the components μ_j^1 and μ_j^2 (the output multipliers), dictated by ranges on transaction processing times. The region Ω_2 would be defined by any restrictions expressing the relative importance of the various inputs pertaining to their impacts on outputs. More will be said regarding such assurance regions later. In general, (6.4) is a constrained version of the original model of Charnes et al. (1978) wherein *linking* constraints that connect output and input bundles are present.

6.2.5 An Alternative Formulation

Model (6.4) can be reduced to a non-ratio format in the usual manner of Charnes and Cooper (1962), yielding:

$$
\begin{aligned}
& e_o^a = \max \mu^1 Y_o^1 + \mu^2 Y_o^2 \\
& \text{subject to:} \\
& v^1 X_o^1 + v^{s_1}\left(\alpha X_o^{s_1}\right) + v^{s_2}((1-\alpha) X_o^{s_2}) + v^2 X_o^2 = 1 \\
& \mu^1 Y_j^1 + \mu^2 Y_j^2 - v^1 X_j^1 - v^{s_1}(\alpha X_j^s) - v^{s_2}(1-\alpha) X_j^s - v^2 X_j^2 \le 0, \forall j \\
& \mu^1 Y_j^1 - v^1 X_j^1 - v^{s_1}(\alpha X_j^2) \le 0 && \forall j \\
& \mu^2 Y_j^2 - v^{s2}((1-\alpha) X_j^2) - v^2 X_j^2 \le 0 && \forall j \\
& 0 \le \alpha_i \le 1, && \forall i \\
& (\mu^1, \mu^2) \varepsilon \Omega_1, (v^1, v^{s_1}, v^{s_2}, v^2) \varepsilon \Omega_2 \\
& \mu_r, v_i \ge \delta, && \forall i, j
\end{aligned}
\qquad (6.5)
$$

Since α_i is a decision variable, this problem is clearly nonlinear. If we make the change of variables $\overline{v}^{s1} = \alpha v^{s1}$ and $\overline{v}^{s2} = (1-\alpha) v^{s2}$, then problem (6.5) reduces to the following form:

$$
\begin{aligned}
&e_o^a = \max \mu^1 Y_o^1 + \mu^2 Y_o^2 \\
&\text{subject to:} \\
&v^1 X_o^1 + \overline{v}^{s_1} X_o^s + \overline{v}^{s_2} X_o^s + v^2 X_o^2 = 1 \\
&\mu^1 Y_j^1 + \mu^2 Y_j^2 - v^1 X_j^1 - \overline{v}^{s1} X_j^s - \overline{v}^{s_2} X_j^s - v^2 X_j^2 \le 0, \quad \forall j \\
&\mu^1 Y_j^1 - v^1 X_j^1 - \overline{v}^{s_1} X_j^2 \le 0 \quad \forall j \\
&\mu^2 Y_j^2 - \overline{v}^{s_2} X_j^2 - v^2 X_j^2 \le 0, \quad \forall j \\
&0 \le \alpha_i \le 1, \quad \forall i \\
&(\mu^1, \mu^2) \in \Omega_1, (v^1, \overline{v}^{s_1}, \overline{v}^{s_2}, v^2) \in \overline{\Omega}_2 \\
&\mu_r^1, \mu_r^2, v_i^1, v_i^2 \ge \delta \\
&\overline{v}_i^{s_1} \ge \alpha_i \delta,\ \overline{v}_i^{s_2} \ge (1 - \alpha_i) \delta
\end{aligned}
\tag{6.6}
$$

The form of $\overline{\Omega}_2$ depends upon how Ω_2 is structured. Clearly, if Ω_2 is the full real space, as is the case when no additional restrictions are imposed on the input multipliers, then (6.6) is a linear programming problem whose solution will immediately yield a solution to the nonlinear model (6.5). In the case that Ω_2 is a proper subset of the real space, defined by restrictions on the input multipliers, then (6.6) may or may not be linear. We consider various types of restrictions on the vectors v, and their impact on the linearity of $\overline{\Omega}_2$, hence model formulation (6.6). Again, we point out that this model is similar to that developed by Beasley (1995) for analyzing the efficiency of universities in terms of teaching and research. In that case the same vector v^s was used for both functions (teaching and research), rather than allowing for different multipliers for vectors on the two components. As a result, Beasley's model does not have an LP equivalent.

6.2.6 Types of Constraints in Ω_2

1. Absolute bounds on the components of $(v^1, v^2, v^{s_1}, v^{s_2})$.

In the case of upper and lower bounds of the form $\delta_1 \le v_i^e \le \delta_2$, where $e = 1, 2, s_1, s_2,$, then $\overline{\Omega}_2$ will consist of linear restrictions since, for example, $\delta_1 \le v_i^{s_1} \le \delta_2$ becomes $\alpha_i \delta_1 \le \overline{v}_i^{s_1} \le \alpha_i \delta_2$.

2. Share of total virtual input occupied by a particular subset of inputs.

Here, we might have constraints of the form

$$\frac{v^{s_1}(\alpha X^s)}{v^{s_1}(\alpha X^s + v^{s_2}((1-\alpha)X^s)} \leq c.$$

Again, such constraints are linear and do not result in nonlinear restrictions in $\overline{\Omega}_2$.

3. Ratio constraints

Restrictions of the cone-ratio variety, see Charnes et al. (1990), may result in nonlinearities in $\overline{\Omega}_2$, depending upon which components of the v-vectors are compared. Specifically, cone-ratio restrictions that do not involve v^{s_1} or v^{s_2} will result in linear constraints in $\overline{\Omega}_2$, for instance the cone-ratio restriction $v^1_{i_1} / v^2_{i_2} \geq c$ can be rewritten as the linear constraint $v^1_{i_1} \geq c v^2_{i_2}$. Ratio constraints on the multipliers of the shared resources *will* render $\overline{\Omega}_2$ nonlinear; for example, restrictions of the form

$$\frac{v^{s_1}_{i_1}}{v^{s_1}_{i_2}} \geq c,$$

are transformed to

$$\frac{\alpha_{i_1} v^{s_1}_{i_1}}{\alpha_{i_2} v^{s_1}_{i_2}} \geq c \frac{\alpha_{i_1}}{\alpha_{i_2}} \text{ or } \frac{\overline{v}^{s_1}_{i_1}}{v^{s_1}_{i_2}} \geq c\,\frac{\alpha_{i_1}}{\alpha_{i_2}},$$

in order to take account of the sharing of resources between sales and service activities.

6.2.7 Special Cases

The extent to which both shared and dedicated resources exist can vary from one situation to another. There can be special circumstances where, for example, there are no dedicated resources and all resources are shared. This does not change the general structure of the constrained DEA model (6.4), nor the requirement that component measures must fall out of the results. One special case is worth noting, namely, when no shared resources are present, and only resources dedicated to the separate components are involved. In this situation, (6.4) is completely separable in the sense that one can derive the individual component measures e^1_o and e^2_o by two separate DEA analyses; one for sales and one for service. The overall aggregate measure e^a_o is then a convex combination of these two measures.

In the following section an application of this multi-component model to a set of bank branches is provided. Due to the presence of ratio constraints of this latter type in the example, the resulting model is nonlinear. In a practical setting with a large number of bank branches to evaluate, solving a quadratic programming problem for each would probably prove to be problematic. A linear relaxation of this nonlinear model is discussed, and outputs from the example are presented. Such a relaxation would prove to be more tractable in the situation where many DMUs are present.

6.3. AN APPLICATION

The model presented herein evolved from an earlier conventional DEA study of branch efficiency in a major Canadian bank. A total of approximately 1300 branches was involved, with the aim of the study being to identify benchmark branches for purposes of establishing cost targets. While data on several hundred different transactions is available from bank records, thirteen of the major ones (some grouped) account for approximately 80% of branch workload, and were used as outputs in the analysis. The only inputs considered in that study were personnel counts. Time studies were conducted previously on a small sample of typical branches, and provide ranges on unit processing times for all transactions. These ranges were the basis for the cone-ratio constraints on output multipliers for the DEA runs performed. One result of the aforementioned study was that members of the set of branches identified as being efficient, were those that were primarily *service* oriented units—specifically those with low levels of activity on the sales side while being very efficient in terms of routine counter transactions. The clear desire of the organization was a methodology that could provide a measure of performance on both components as well as an overall efficiency score. In this way one can identify not only those branches that are underperforming, but also the component that is weakest. The model discussed in Section 2 was applied to a dataset of 20 branches out of the full set of bank branches. These were all chosen from one district. For purposes of illustration only, a subset of transaction types was chosen as outputs, and only personnel counts were used as inputs. The chosen input- and output measures used are summarized in Table 6-1.

Table 6-1. Input- and Output measures used in an application of the model

Inputs		Outputs	
FSE	# service staff	MDP	# counter level deposits
FSA	# sales staff	MTR	# transfers between accounts
FSU	# support staff	RSP	# retirement savings plan openings
FOT	# other staff	MOR	# mortgage accounts opened

The relevant data for a one year period is displayed in Table 6-2. To provide for a realistic picture of branch performance, a number of restrictions were imposed:

Table 6-2. Branch Data for a selection of 20 bank branches

	service outputs		sales outputs		inputs		shared inputs	
DMU	MDP	MTR	RSP	MOR	FSE	FSA	FSU	FOT
01	2.873	1.498	03.6	04.2	0.455	0.492	0.17	0.73
02	3.093	1.226	05.9	09.7	0.942	0.661	1.88	1.00
03	1.857	0.865	03.7	04.9	0.510	0.293	0.47	1.01
04	8.532	3.290	04.8	12.2	1.239	0.916	1.13	0.10
05	4.304	1.777	07.9	16.8	1.015	0.724	4.48	0.12
06	4.340	0.110	00.5	00.9	0.883	1.474	3.61	0.33
07	4.640	1.493	08.7	05.2	0.594	0.320	2.86	0.21
08	6.821	3.243	07.4	11.0	0.815	0.669	2.99	0.16
09	4.709	2.599	06.5	06.3	0.862	0.670	0.92	1.21
10	0.015	0.037	00.6	02.9	0.000	0.060	5.45	1.55
11	8.532	4.332	09.7	07.2	0.972	1.216	0.12	0.14
12	5.312	2.718	03.5	03.5	0.035	1.007	0.42	0.31
13	3.643	2.115	08.4	06.4	1.317	0.550	2.59	0.17
14	4.878	3.010	05.9	06.0	0.610	0.939	0.54	0.12
15	4.109	1.993	06.0	06.2	0.511	0.659	1.96	0.01
16	4.950	2.950	05.3	04.7	0.719	0.602	1.17	0.49
17	6.389	2.415	12.3	07.8	1.485	0.689	5.03	0.26
18	2.939	1.377	09.0	04.3	0.528	0.436	0.39	0.13
19	6.184	1.975	02.7	04.3	0.743	0.546	0.83	0.56
20	3.053	0.951	01.0	03.2	0.508	0.395	1.44	1.25

Type 1: Ratio constraints on multipliers

Ratio constraints of the form $a \le \mu_{r_1} / \mu_{r_2} \le b$ on output multipliers were imposed to reflect processing times. Ratio constraints on the shared input multipliers were applied to reflect the relative importance of the two inputs (support and other staff) that are split between sales and service.

Type 2: Limitations on α_i

It is generally the case that some bounds need to be imposed on the fraction α_i of shared resource i being allocated to service activities. For illustrative purposes the range $1/3 \le \alpha_i \le 2/3$ was chosen.

Type 3: Constraints on the ratios of total service inputs to total inputs.

Here constraints are imposed to restrict the portion of virtual inputs being allocated to the service component. Recalling the definition of β_j in Property 1, restrictions were imposed on the range over which β_j could vary. For present purposes the limits $1/3 \le \beta_j \le 2/3$ were applied. While the same limits were used for all branches j in the example herein, it may be the case that different ranges would apply to different classes of branches. Large urban branches may allocate different mixes of resources to sales than small or mid-size branches.

6.3.1 Model Relaxation

The model presented in the previous section is nonlinear in the presence of ratio constraints (Type 1) on shared input multipliers. Specifically, when we impose constraints $\alpha \leq v_1^{s_1} / v_2^{s_1} \leq b$, these take the form

$$a\frac{\alpha_1}{\alpha_2} \leq \frac{v_1^{s_1}\alpha_1}{v_2^{s_1}\alpha_2} \leq b\frac{\alpha_1}{\alpha_2}$$

in the presence of the transformation discussed in Section 3. To render the model more tractable, various *linear relaxations* are possible. One approach attempted was iterative. Specifically, in the first stage all α_i are assumed to be equal for any given branch (i.e., $\alpha_i = d$, a single variable), and the resulting linear problem was solved to determine a starting solution. This yields an optimal solution $(\mu^*_{(1)}, v^*_{(1)}, \alpha^*_{(1)})$. Fixing $\mu = \mu^*_{(1)}$ and $v = v^*_{(1)}$, the second stage derives a best set of $\alpha^*_i(2)$ relative to the constants $\mu^*_{(1)}$ and $v^*_{(1)}$. In subsequent stages one alternately fixes either $\alpha^*(n)$ or the pair $(\mu^*_{(n)}, v^*_{(n)})$, and optimizes (6.4) on the other. One of the difficulties encountered with this method was that many iterations were required in order to converge to a solution that was reasonably close to the optimum.

An alternative and somewhat more practical method was investigated. This amounted to choosing a grid of points in each α_i range. In the present case, each of the two α_i ranged from 0.25 to 0.75 and the grid of 5 values 0.25, 0.35, 0.45, 0.65, 0.75 was used. Recall that α_1 is the percent of "support staff" allocated to service transactions and that α_2 represents the split of "other staff". This resulted in 5×5=25 different combinations for (α_1, α_2).

Given the relatively small sample of DMUs in this particular example (20 DMUs), the problem can easily be treated directly in its nonlinear form, and was solved using a standard spreadsheet solver.

6.3.2 Results

A proper evaluation of data such as that in Table 6-2 is complicated by the fact that the sales component is a two-level process as discussed earlier. The ranges for average processing times, as reflected in the cone-ratio constraints imposed upon the output multipliers, pertain only to the second of these two levels, namely the *transaction* part of sales. These average times do not account for the *level of effort* required to transact the sale. This effort would involve activities such as interaction with customers, review of portfolios, etc. To compensate for the *understated* values of the μ_j components, one must either scale up these values, or adjust (downward) the resources (inputs) allotted to the sales component. The latter option becomes problematic in that the portion of sales resources not allocated to the transaction part of sales is left as unassigned inputs (i.e., they appear to not contribute to any of the outputs). In the present situation, the former option of scaling up the sales output multipliers was chosen. The scaling factor γ, defined as the ratio of the "Total Sales effort" to the "Transaction effort" was based on an estimate provided by the organization. The ranges provided for μ_j, namely $a \le \mu_j \le b$, were replaced by scaled ranges $\gamma a \le \mu_j \le \gamma b$. The resulting aggregate, service and sales efficiency scores are displayed in Table 6.3. It is noted that only one of the branches (#11), is efficient in the aggregate sense, that is in both sales and service. Clearly, branches may be efficient in one component only, such as is the case for branches #12 and #18. The respective α_1 and α_2 values are also shown.

Table 6-3. Efficiency Scores and Optimal Split of Shared Resources

DMU	Aggregate e_k^a	Service e_k^1	Sales e_k^2	α_1	α_2
01	0.47972	0.52172	0.45354	0.72676	0.75000
02	0.40499	0.17158	0.52749	0.75000	0.25000
03	0.41946	0.23162	0.50145	0.75000	0.25000
04	0.74913	0.51905	0.91297	0.75000	0.64929
05	0.54472	0.17250	0.54472	0.75000	0.75000
06	0.14925	0.17663	0.03273	0.75000	0.66891
07	0.47257	0.28014	0.55697	0.75000	0.75000
08	0.58236	0.38787	0.70302	0.36427	0.29968
09	0.41178	0.36773	0.43157	0.25000	0.55019
10	0.07307	0.00570	0.09894	0.26281	0.68108
11	1	1	1	0.75000	0.66891
12	0.57384	1	0.29015	0.75000	0.74959
13	0.40464	0.17685	0.53991	0.53334	0.75000
14	0.70675	0.71811	0.70001	0.53334	0.75000
15	0.49252	0.36720	0.55537	0.75000	0.75000
16	0.44784	0.46087	0.43869	0.25000	0.54547
17	0.36581	0.19350	0.45445	0.25000	0.72687
18	0.85924	0.46010	1	0.25000	0.72687
19	0.49243	0.52389	0.37181	0.72682	0.72188
20	0.21444	0.26296	0.18235	0.72676	0.72224

6.4. MEASURING MULTI COMPONENT EFFICIENCY – AN ADDITIVE MODEL

6.4.1 Addressing Some Shortcomings

The model described above, when applied within the organization, did help to point to areas where inefficiency existed within branches, and aided in setting targets for improvements. Two suggestions from management for enhancement of performance measurement arose from this application.

6.4.1.1 Non-Volume Related Activities

The first issue has to do with the characterization of those activities surrounding the sales function. The sales function within the bank environment can be viewed as consisting of two sets of activities. The first set, and those examined in the previous sections, would be classified as *volume-related* activities. These activities consist of those tasks linked directly to sales products, *after* the decision to purchase has been made. These would include the filing of documents, preparation of certificates, etc. Such tasks are characterized by known time estimates, arrived at in the same manner as is the case for service transactions.

The second set, the *non-volume-related* activities, may not be directly linked to any specific product. Such activities would include responding to customer queries, routine tasks such as reproduction of forms, reviewing customer portfolios, carrying out computer searches, and so on. Support costs for print materials, computer expenses, etc. would, as well, fall into this category.

6.4.1.2 Providing a Fair Balance Between Sales and Service Performance Measures

The model of the previous section, because of the form of the objective function, will often produce component measures e_j^1 and e_j^2 that differ from each other in an unreasonable way. Essentially, the model, in setting out to maximize the aggregate score e_j^a will do so by maximizing one of the two component measures *at the expense of the other.* A suggestion raised by management was to attempt to derive measures with the idea of showing both sales and service performance in the best light. To address the above two concerns, an additive form of the DEA model was adopted.

6.4.2 The General Additive Model

In the next subsection we develop a dual-component DEA model for evaluating both sales and transaction functions within bank branches. For purposes of that development, the Pareto-Koopmans, or additive model structure is exploited. While the additive model is seldom the structure of choice in most DEA analyses (one generally utilizes one of the radial models), it is demonstrated that it's structure is, in fact, a general framework containing the radial models as special cases. Specifically, any of the standard models are obtainable by way of constrained versions of the additive model. For development purposes herein, it is convenient to approach the standard models from this angle, rather than in the more conventional way.

It is instructive to examine both dual and primal forms of the additive model:

The Dual

$$\min \sum_i v_i x_{i_o} - \sum_r \mu_r y_{r_o} - \mu_o \tag{6.7a}$$

subject to:

$$\sum_i v_i x_{ij} - \sum_r \mu_r y_{rj} - \mu_o \geq 0, \forall j \tag{6.7b}$$

$$\mu_r \geq 1/y_{r_o}, \forall r \tag{6.7c}$$

$$v_i \geq 1/x_{i_o}, \forall i \tag{6.7d}$$

It is noted that we have chosen lower bounds on the multipliers (6.1c) and (6.7d) that are DMU-specific. This is usually referred to as the *units invariant* form of the model. The "dual" of (6.7a) is the model:

The primal

$$\max \sum_i (s_i^1 / x_{i_o}) + \sum_r s_r^2 / y_{r_o} \tag{6.8a}$$

subject to:

$$\sum_j \lambda_i x_{ij} + s_i^1 \leq x_{i_o}, \ \forall i, \tag{6.8b}$$

$$\sum_j \lambda_j y_{rj} - s_r^2 \geq y_{ro}, \quad \forall r, \tag{6.8c}$$

$$\sum_j \lambda_j = 1, \tag{6.8d}$$

$$s_i^1, s_r^2, \lambda_j \geq 0, \ \forall i, r, j. \tag{6.8e}$$

If we adopt the notation

$$\theta_i = 1 - s_i^1/x_{i_o}, \phi_r = 1 + s_r^2/y_{r_o} \tag{6.9}$$

and let $\bar{\theta}_i = 1 - \theta_i, \bar{\phi}_r = \phi_r - 1$, model (6.8a)-(6.8e) becomes

$$\max \sum_i \bar{\theta}_i + \sum_r \bar{\phi}_r \tag{6.10a}$$

subject to:

$$\sum_j \lambda_i x_{ij} + \bar{\theta}_i x_{i_o} \leq x_{i_o}, \forall i \tag{6.10b}$$

$$\sum_j \lambda_j y_{rj} - \bar{\phi}_r y_{ro} \geq y_{r_o}, \forall r \tag{6.10c}$$

$$\sum_j \lambda_j = 1 \tag{6.10d}$$

$$\bar{\theta}_i, \bar{\phi}_r, \lambda_j \geq 0, \forall i, r, j \tag{6.10e}$$

This format is a particularly convenient way to view the additive model, as it exhibits an immediate connection to other models. This form is related to the "Russell Measure" as discussed in Fare and Lovell (1978). There, the objective function takes the form

$$\min R = \left[\sum_i \theta_i + \sum_r (1/\phi_r) \right] / (I + R),$$

where I,R are the numbers of inputs and outputs, respectively. Cooper et al. (1999) discuss several variations on the additive model, as does Thrall (1996).

It is immediately clear that one can adopt a purely *input oriented* variation on the additive model concept, by setting $\bar{\phi}_r = 0$ for all r, and replacing constraints (6.10b) and (6.10c) by

$$\sum_j \lambda_j x_{ij} + \bar{\theta}_i x_{io} \leq x_{io} \tag{6.11a}$$

$$\sum_j \lambda_j y_{rj} \geq y_{ro} \tag{6.11b}$$

This type of structure is discussed in Zieschang (1984). In the section to follow we focus attention on the input oriented model. Futhermore, if we restrict the $\bar{\theta}_i$ further by requiring that they all be equal, then we have a structure equivalent to the standard input oriented *radial* model of Charnes, Cooper and Rhodes (1978) (or at least Banker et al. (1984)).

In the case that the input oriented approach is to be taken, in which case (6.11a) and (6.11b) replace (6.10b) and (6.10c) in the primal problem (6.10a), the equivalent modification to the dual problem (6.7a) is to replace the lower bound on μ_r (constraint (6.7c) by $\mu_r \geq 0$. As with the Russell measure, an appropriate measure of performance in the input oriented additive model is

$$R_I = \sum_{i=1}^{I}(1-\bar{\theta}_i)/I = \sum_{i=1}^{I}\theta_i/I. \tag{6.12}$$

It is noted that in the restricted case where $\theta_i = \theta$ for all i (the BCC radial model), $R_I = \theta$. In any event, it will be the case that $0 \le R_I \le 1$, with $R_I = 1$ if all $\bar{\theta}_i = 0$; for example, in this case the pair (Y^0, X^0) is on the frontier or an extension.

Stated formally then, the pure input version of (6.10a)-(6.10e) is:

$$\begin{aligned}
&\max \sum_i \bar{\theta}_i / I \\
&\text{subject to:} \\
&\sum_j \lambda_j x_{ij} + \bar{\theta}_i x_{i_o} \le x_{i_o}, \ \forall i \\
&\sum_j \lambda_j y_{rj} \ge y_{ro}, \ \forall r \\
&\sum_j \lambda_j = 1 \\
&\bar{\theta}_i, \lambda_j \ge 0
\end{aligned} \tag{6.13}$$

Thus, the additive model can be viewed as a flexible mechanism for capturing different aspects of efficiency. Admittedly, restricted versions of the model can fail to be comprehensive in the sense discussed by Cooper et al. (1999). Obviously, it will be true that restricting attention to the input side of the problem, for example, can mean that improper envelopment can occur, as is well known in the radial models.

6.4.3 An Additive Model for Sales and Service Components

The notation of the previous section will be used in the current model, but with the one addition, namely, to use two output multipliers μ^{21} for the per unit processing times for volume-related and μ^{22} for non-volume related portions of the sales outputs Y_j^2. We also chose in this second analysis to use the VRS DEA model, hence defined output variables μ_o^1 and μ_o^2 for service and sales components.

An alternative to optimizing the aggregate efficiency measure as in the previous sections, is to attempt to optimize, in some manner, both the service measure

$$e_{1_o} = \nu^1 X_o^1 + \nu^{s1}(\alpha X_o^s) - \mu^1 Y_o^1 - \mu_o^1 \tag{6.14}$$

and sales measure[1]

$$e_{2_o} = \nu^2 X_o^2 + \nu^{s2}((1-\alpha)X_o^s) - \mu^{21}Y_o^2 - \mu^{22}Y_o^2 - \mu_o^2. \quad (6.15)$$

One approach is to minimize the maximum inefficiency, for example, we solve the goal programming problem.

$$\begin{aligned} &\min \ d \\ &\text{subject to:} \\ &e_{1_o} \le d,\ e_{2_o} \le d \\ &S_e I_j - S_e O_j \ge 0,\ S_a I_j - S_a O_j \ge 0,\ \forall j. \end{aligned} \quad (6.16)$$

In attempting to reduce the maximum inefficiency (d), the model has the tendency to equalize the sales and service performance measures if feasibility permits. In some respects this could be justified insofar as one can argue that a branch will, or should, give equal importance to all components of its business. It must be pointed out that additional restrictions may be imposed on the multipliers in (6.16) (e.g., assurance regions as per Thompson et al. (1990)). For example, the components of μ^1 would be related to one another through limits arising from branch time studies. For model development purposes in this section, however, we avoid applying specific additional restrictions. This permits us to obtain primal and dual efficiency measurement models, not tied to application-specific situations. The inclusion of these in the models is examined in the next section dealing with the application of the tools in a specific setting.

[1]In the context of the VRS structure, we let μ_o^1, μ_o^2 denote service and sales variables, respectively.

Formally, the dual form of the proposed model is given by (6.17).

$$
\begin{aligned}
&\min\ d\\
&\text{subject to:}\\
&-v^1X_o^1 - v^{s1}(\alpha X_o^s) + \mu^1 Y_o^1 + \mu_o^1 + d \geq 0\\
&v^1X_j^1 + v^{s1}(\alpha X_j^s) - \mu^1 Y_j^1 - \mu_o^1 \geq 0,\ \forall j,\\
&-v^2X_o^2 - v^{s2}((1-\alpha)X_o^s) + \mu^{21}Y_o^2\\
&+\mu^{22}Y_o^2 + \mu_o^2 + d \geq 0\\
&v^2X_j^2 + v^{s2}((1-\alpha)X_j^s) - \mu^{21}Y_j^2\\
&-\mu^{22}Y_j^2 - \mu_o^2 \geq 0,\ \forall j, \qquad\qquad (6.17)\\
&v_i^1 \geq 1/(x_{i_o}^1 . |I_1|),\ \forall i \in I_1,\\
&v_i^2 \geq 1/(x_{i_o}^2 . |I_2|), \forall i \in I_2,\\
&v_i^{s1} \geq 1/(x_{i_o}^s . |I_s|),\quad \forall i \in I_s,\\
&v_i^{s2} \geq 1/(x_{i_o}^s . |I_s|),\quad \forall i \in I_s,\\
&\mu_r^1 \geq 0,\quad \forall r \in R,\\
&\mu_r^{21} \geq 0,\quad \forall r \in R,\\
&\mu_r^{22} \geq 0,\quad \forall r \in R,
\end{aligned}
$$

Note that we have introduced the lower bounds $1/x_o^1 . |I_1|$, etc., to force $0 \leq d \leq 1$. Here, $|I_1|$ denotes the cardinality of the input set I_1.

To deal with the nonlinearity created by the products αv^{s1} and $(1-\alpha)v^{s2}$, introduce the change of variables $\overline{v}^{s1} = \alpha v^{s1}$, and $\overline{v}^{s2} = (1-\alpha)v^{s2}$.

Then, replace the two constraints $v_i^{s1} \geq 1/(x_{io}^s . |I_s|)$ and $v_j^{s2} \geq 1/(x_{io}^s\ |I_s|)$ by $\alpha v_i^{s1} \geq \alpha 1/(x_{io}^s . |I_s|)$ and $(1-\alpha)\ v_j^{s2} \geq 1/(x_{io}^s\ |I_s|)$ - $\alpha 1/(x_{io}^s\ |I_s|)$.

It is generally the case that constraints will be imposed on the α_i; specifically, the percent of any resource that can be allocated to the service component will be required to be within some interval, namely

$$L_i^1 \leq \alpha_i \leq L_i^2.$$

Model (6.17) can now be rewritten in the form:

$$
\begin{aligned}
e_p = \max & \left\{ \sum_{i \in I_1} s_i^1 / (x_{io}^s \mid I_1 \mid) + \sum_{i \in I_2} s_i^2 / (x_{io}^2 \mid I_2 \mid) \right. \\
& \left. + \sum_{i \in I_s} \left[s_i^{s_2} / (x_{io}^s \mid I_s \mid) + L_i^1 \gamma_i^1 - L_i^2 \gamma_i^2 \right] \right\} \\
& \text{subject to:} \\
& \sum_k \lambda_k^1 x_{ik}^1 - \lambda_{n+1}^1 x_{io}^1 + s_i^1 \leq 0, \ i \in I_1, \\
& \sum_k \lambda_k^1 x_{ik}^s - \lambda_{n+1}^1 x_{io}^s + s_i^{s1} \leq 0, \ i \in I_s, \\
& \sum_k \lambda_k^2 x_{ik}^2 - \lambda_{n+1}^2 x_{io}^2 + s_i^2 \leq 0, \ i \in I_2, \\
& \sum_k \lambda_k^2 x_{ik}^s - \lambda_{n+1}^2 x_{io}^s + s_i^{s2} \leq 0, \ i \in I_s, \\
& -\sum_k \lambda_k^1 y_{rk}^1 - \lambda_{n+1}^1 y_{ro}^1 \leq 0, \ r \in R_1, \\
& -\sum_k \lambda_k^2 y_{rk}^2 + \lambda_{n+1}^2 y_{ro}^2 \leq 0, \ r \in R_2, \\
& \lambda_{n+1}^1 + \lambda_{n+1}^2 = 1 \\
& -s_i^{s1} / (x_{io}^s . \mid I_s \mid) + s_i^{s2} / (x_{io}^s . \mid I_s \mid) \\
& + \gamma_i^1 - \gamma_i^2 \leq 0, \\
& \gamma_i^1, \gamma_i^2, \lambda_k^1, \lambda_k^2, s_i^1, s_i^2 \geq 0.
\end{aligned}
\tag{6.18}
$$

The dual of this problem is given by

$$e_p = \max\left\{\sum_{i\in I_1} s_i^1/(x_{io}^s \mid I_1 \mid) + \sum_{i\in I_2} s_i^2/(x_{io}^2 \mid I_2 \mid) + \sum_{i\in I_s}\left[s_i^{s_2}/(x_{io}^s \mid I_s \mid) + L_i^1\gamma_i^1 - L_i^2\gamma_i^2\right]\right\}$$

$$\begin{aligned}
&\text{subject to:}\\
&\sum_j \lambda_j^1 x_{ij}^1 - \lambda_{n+1}^1 x_{io}^1 + s_i^1 \le 0,\ i \in I_1,\\
&\sum_j \lambda_j^1 x_{ij}^s - \lambda_{n+1}^1 x_{io}^s + s_i^{s1} \le 0,\ i \in I_s,\\
&\sum_j \lambda_j^2 x_{ij}^2 - \lambda_{n+1}^2 x_{io}^2 + s_i^2 \le 0,\ i \in I_2,\\
&\sum_j \lambda_j^2 x_{ij}^s - \lambda_{n+1}^2 x_{io}^s + s_i^{s2} \le 0,\ i \in I_s,\\
&-\sum_j \lambda_j^1 y_{rj}^1 - \lambda_{n+1}^1 y_{ro}^1 \le 0,\ r \in R_1,\\
&-\sum_j \lambda_j^2 y_{rj}^2 + \lambda_{n+1}^2 y_{ro}^2 \le 0,\ r \in R_2,\\
&\lambda_{n+1}^1 + \lambda_{n+1}^2 = 1\\
&-s_i^{s1}/(x_{io}^s.\mid I_s \mid) + s_i^{s2}/(x_{io}^s.\mid I_s \mid)\\
&+\gamma_i^1 - \gamma_i^2 \le 0,\\
&\gamma_i^1, \gamma_i^2, \lambda_j^1, \lambda_j^2, s_i^1, s_i^2 \ge 0.
\end{aligned} \tag{6.19}$$

Letting $\overline{\theta}_i^1 = s_i^1/x_{io}^1, \overline{\theta}_i^2 = s_i^2/x_{io}^2, \overline{\theta}_i^{s1} = s_i^s/x_{io}^s, \overline{\theta}_i^{s2} = s_i^{s2}/x_{io}^s$, problem (6.19) can be written as

$$
\begin{aligned}
e_p = \max\Bigg\{ & \sum_{i\in I_1} \bar{\theta}_i^1 / |I_1| + \sum_{i\in I_2} \bar{\theta}_i^2 / |I_2| \\
& + \sum_{i\in I_s} \left[\bar{\theta}_i^{s_2} / |I_s|) + L_i^1\gamma_i^1 - L_i^2\gamma_i^2 \right] \Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
& \text{subject to:} \\
& \sum_j \lambda_j^1 x_{ij}^1 - \lambda_{n+1}^1 x_{io}^1 + \bar{\theta}_i^1 x_{io}^1 \le 0, \ i \in I_1, \\
& \sum_j \lambda_j^1 x_{ij}^s - \lambda_{n+1}^1 x_{io}^s + \bar{\theta}_i^{s1} x_{io}^s \le 0, \ i \in I_s, \\
& \sum_j \lambda_j^2 x_{ij}^2 - \lambda_{n+1}^2 x_{io}^2 + \bar{\theta}_i^2 x_{io}^2 \le 0, \ i \in I_2, \\
& \sum_j \lambda_j^2 x_{ij}^s - \lambda_{n+1}^2 x_{io}^s + \bar{\theta}_i^{s2} x_{io}^s \le 0, \ i \in I_s, - \\
& \sum_j \lambda_j^1 y_{rj}^1 - \lambda_{n+1}^1 y_{ro}^1 \le 0, \ r \in R_1, \\
& -\sum_j \lambda_j^2 y_{rj}^2 + \lambda_{n+1} y_{ro}^2 \le 0, \ r \in R_2, \\
& \lambda_{n+1}^1 + \lambda_{n+1}^2 = 1, \\
& -\bar{\theta}_i^{s1} / |I_s| + \bar{\theta}_i^{s2} / |I_s| + \gamma_i^1 - \gamma_i^2 \le 0, \\
& \gamma_i^1, \gamma_i^2, \lambda_j^1, \lambda_j^2, \bar{\theta}_i^1, \bar{\theta}_i^2 \ge 0.
\end{aligned}
\qquad (6.20)
$$

It can be seen that (6.20) is a direct generalization of (6.13). The equivalent of the R_1 measure given in (6.12) is $\bar{e}_p = 1 - e_p$.

It must be noted, of course that $e_p = e_d$ (the objective function value of (6.18)) is the maximum of the two components e_{1_0}, e_{2_0}, as per (6.2) and (6.3). the separate sales and service measures would fall out as part of the analysis.

6.5. APPLICATION TO BANK BRANCHES

To demonstrate the application of the additive structure, we again examine data on a sample of branches, with somewhat different outputs. Data on twenty branches is displayed in Table 6.4. The outputs chosen were:

Service:
TOTEMU – total number of menu account transactions

VISA – number of Visa cash advances
CAD – number of commercial deposit transactions

Sales:
RSP – number of RSP account openings
MORT – number of mortgages transacted
BPL – number of variable rate consumer loans transacted

In the current example, inputs were restricted to personnel only. We have not included other operating expenses such as computers, rent, etc. Specifically, the inputs were:

Service: FSE – total number of full time equivalent service staff
Sales: FSA – total number of full time equivalent sales staff
Shared: FSU – total number of full time equivalent support staff
FST – total number of full time equivalent "other" staff.

Table 6-4. Sales and Service Outputs and Inputs

Transit #	TOTMENU	VISA	CAD	RSP	MORT	BPL	FSE	FSA	FSU	FOT
1	51803	2973	190522	421	567	101	55.25	98.15	48.07	59.55
2	10477	710	49898	75	172	13	14.73	32.39	17.15	7.24
3	11195	431	39523	68	73	19	9.6	14.22	8.55	2.15
4	6480	422	30713	49	48	15	10.48	9.05	5.15	1
5	37695	921	26922	210	128	144	27.9	17.22	5.02	1.08
6	9211	362	43056	120	127	57	9.17	15.79	2.44	1
7	16483	529	13123	74	150	9	10.53	12.49	2.19	1.41
8	456	20	10127	6	29	15	10.6	10	5.45	1.55
9	5985	382	21945	32	28	19	8.71	8.02	3.94	1
10	8682	351	11010	84	78	52	7.05	11.25	2.36	0.96
11	5287	182	16474	59	97	15	6.61	9.42	1.88	1
12	18292	171	18014	104	84	443	7.3	9.33	1.79	1.05
13	5669	264	11303	36	62	144	3.96	3.92	1	0.8
14	9656	332	6745	65	63	14	6.7	8.62	0.92	1.21
15	18566	308	76174	134	80	14	10.29	10.94	4.87	1.03
16	39430	500	60832	199	199	200	25.07	20.55	4.69	0.83
17	11601	423	36692	73	137	107	12.25	10.91	3.88	0.96
18	8030	406	19598	62	86	50	9	13.35	3.16	0.97
19	16991	658	21334	91	111	78	12.7	18.02	2.11	1.76
20	10473	463	51225	132	71	39	18.15	18.65	6.92	1

As discussed earlier, in applying the models described herein, attention was paid to multiplier restrictions that reflect the relative weights to be placed on the various outputs. Specifically, the components of μ^1 and μ^{21} have been constrained in a ratio sense to obey time limits on branch transactions. For example, the specified time interval for a commercial deposit transaction is (in minutes) (1.2, 3.6); that for a VISA cash advance is (0.8, 2.5). To reflect these limits in the multipliers μ_2^1 and μ_3^1, we require $\frac{0.8}{3.6} \leq \mu_2^1/\mu_3^1 \leq \frac{2.5}{1.2}$. Similar restrictions have been applied to the components of μ^{21} to accommodate the time limits on the transaction portion of sales outputs.

Table 6-5. Results from Model

Var1		2	3	4	5	6	7 8	8
α_1	0.742798	0.75	0.25	0.25	0.25	0.371086	0.746744	0.25
α_2	0.5	0.5	0.278917	0.5	0.25	0.5	0.5	0.5
d	70.14%	85.13%	40.00%	80.24%	0.00%	23.59%	94.27%	34.77%
ES	29.86%	14.87%	60.00%	19.76%	100.00%	76.41%	5.73%	65.23%
ET	29.86%	14.87%	60.00%	19.76%	100.00%	76.41%	5.73%	65.23%
Var	9	10	11	12	13	14	15	16
α_1	0.413508	0.507633	0.25	0.25	0.25	0.25	0.25	0.25
α_2	0.5	0.75	0.25	0.57965	0.417598	0.75	0.75	0.675252
d	29.79%	23.65%	66.80%	0.00%	0.00%	0.00%	0.00%	0.00%
ES	70.21%	76.35%	48.50%	100%	100%	100%	100%	88.01%
ET	70.21%	76.35%	33.20%	100%	100%	100%	100%	100%
Var	17	18	19	20	Average	SDev.		
α_1	0.25	0.75	0.75	0.25	40.16%	21.62%		
α_2	0.438533	0.75	0.75	0.75	54.45%	17.12%		
d	16.94%	68.09%	62.51%	65.80%	38.09%	33.18%		
ES	89.71%	31.92%	37.49%	34.20%	62.41%	32.30%		
ET	83.06%	31.92%	37.49%	34.20%	61.91%	33.18%		

No such detailed information was obtainable on the non-volume portion of the sales component. From interviews with branch consultants, it has been estimated that 30% to 50% of the sales effort lies with the non-volume activity, and the remainder is the transaction or volume-related work. In general, this would imply that $\frac{3}{7} \leq \mu^{22}y_j^2/\mu^{21}y_j^2 \leq 1$ for each branch k. To simplify matters, we choose here to take a more restricted view, and constrain the ratio *for each product* i to be in this range. Specifically, $\frac{3}{7} \leq \mu_i^{22}y_{ij}^2/\mu_i^{21}y_{ij}^2 \leq 1$, implying that $\frac{3}{7} \leq \mu_i^{22}/\mu_i^{21} \leq 1$.

Table 6-4 displays the data on all inputs and outputs for a sample of 20 branches of the bank. The result from applying model (6.18), (augmented by the multiplier restrictions discussed above), are shown in Table 6-5. Recall that d represents the maximum inefficiency associated with the two components (sales and service). The corresponding efficiency measures e_s (sales) and e_T (service) are displayed. It is noted that $d = 1 - \min\{e_T, e_S\}$. As noted earlier, this model tends to force e_T and e_S together, and in a large percentage of the cases, the two measures are equal.

We have not directly addressed the issue of an aggregate measure of efficiency which should be some combination of the two separate measures. Arguably, this aggregate measure e_A should be some average of the component scores. A reasonable candidate for this might be of the form $e_A = \beta e_T + (1-\beta)e_S$ where β is the proportion of total resources consumed by the service component (dedicated service inputs together with shared inputs).

In the application of model (6.18), the splitting variables α_1 and α_2 have each been restricted to the range $.25 \leq \alpha \leq .75$. This range would need to be established by branch consultants in much the same manner that ranges on output multiplies might be set by way of time study estimates.

6.6. CONCLUSIONS

This chapter has examined model structures for dealing with multi-component efficiency measurement in a banking environment. The conventional DEA approach, as applied in bank related studies, has tended to concentrate on a single measure of performance for the DMU. Very often, however, there are multiple components or sub units within the DMU whose individual performance is required. The model provided herein provides a mechanism for developing multi-component measures.

REFERENCES

1. Banker, R.D., A. Charnes and W.W. Cooper (1984), Some models for technical and scale efficiencies in Data Envelopment Analysis, *Management Science* 30, 1078-1092.
2. Beasley, J.E. (1995), Determining teaching and research efficiencies, *Journal of the Operational Research Society* 46, 441-52.
3. Beasley, J.E. and Y.-H. B. Wong (1990), Restricting weight flexibility in data envelopment analysis, *Journal of the Operational Research Society* 41, 829-835.
4. Charnes, A., W.W. Cooper and E.L. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2, 429-44.

5. Charnes, A., W.W. Cooper and Z.M. Huang and D.B. Sun (1990), Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks, *Journal of Econometrics* 46, 73-91.
6. Charnes, A., W.W. Cooper and E.L. Rhodes (1978), Measuring the efficiency of Decision Making Units, *European Journal of Operational Research* 2, 429-444.
7. Cook, W.D., and M. Hababou (2001), Sales performance measurement in bank branches, *OMEGA* 29, 299-307.
8. Cook, W.D., M. Hababou and H. Tuenter (2000), Multicomponent efficiency measurement and shared inputs in data envelopment analysis: An application to sales and service performance in bank branches, *Journal of Productivity Analysis* 14, 209-24.
9. Cooper, W., K. Park and J. Pastor (1999), RAM: A range adjusted measure of inefficiency for use with additive models, and relations to other models and measures in DEA, *Journal of Productivity Analysis* 11, 5-42.
10. Fare, R., C.A.K. Lovell (1978), Measuring the technical efficiency of production, *Journal of Economic Theory* 19, 150-62.
11. Fare, R. And S. Grosskopf (1996), Productivity and intermediate products: a frontier approach, *Economic Letters* 50, 65-70.
12. Fare, R., and D. Primont (1984), Efficiency measures for multiplant firms, *Operations Research Letters* 3, 257-260.
13. Oral, M. And R. Yolalan (1990), An empirical study on measuring operational efficiency and profitability of bank branches, *European Journal of Operational Research* 46, 282-294.
14. Schaffnit, C., D. Rosen and J.C. Paradi (1997), Best practice analysis of bank branches: an application of data envelopment analysis in a large Canadian bank, *European Journal of Operational Research*, 98, 269-89.
15. Sherman, H.D. and F. Gold (1985), Bank branch operating efficiency: evaluation with Data Envelopment Analysis, *Journal of Banking and Finance* 9, 297-315.
16. Sherman, H.D., and G. Ladino (1995), Managing bank productivity using data envelopment analysis (DEA), *Interfaces* 25, 60-73.
17. Thompson, R.G., L.N. Langemeier and C.T.Lee and E. Lee and R.M. Thrall (1990), The role of multiplier bounds in efficiency analysis with application to Kansas farming, *Journal of Econometrics* 46, 93-108.
18. Thrall, R.M. (1996), The lack of invariance of optimal dual solutions under translations, *Annals of Operations Research* 66, 103-8.
19. Zieshang, K. (1984), An extended Farrell efficiency measure, *Journal of Economic Theory* 33:387-96.

This chapter is based upon Cook, W.D., M. Hababou and H. Tuenter, 2000, Multicomponent efficiency measurement and shared inputs in Data Envelopment Analysis: An application to sales and service performance in bank branches, *Journal of Productivity Analysis*, 14(3), 14:209-24, with permission from Kluwer Academic Publishers, and Cook, W.D., and M. Hababou, (2001), Sales performance measurement in bank branches, *OMEGA*, 29:299-307, with permissions from Elsevier Science.

Chapter 7

DEA AND MULTICRITERIA DECISION MODELING

7.1. INTRODUCTION

Many real world problems involve evaluating a set of alternatives or choices when multiple criteria need to be considered. The general DEA model is an example of a special type of multiple criteria decision model (MCDM) framework, wherein both outputs and inputs are present. A number of the chapters that follow examine various types of problems, all of which fall within the generic MCDM framework. Typically, such problems can involve both quantitative and qualitative data. In the current chapter we examine the usage of the DEA methodology to tackle such problems.

A linear composite index is a simple, straightforward and intuitive means for evaluating alternatives in the presence of multiple criteria. In notational terms, if v_{ij} represents the score or evaluation of alternative i with respect to criterion j and W_j is the weight or importance associated with criterion j, then the linear composite index is given by $R_i = \sum_j W_j v_{ij}$. If, for example, the alternatives are proposed capital projects to be undertaken by an organization, then the $\{R_i\}_i$ provide a basis for judging which projects should be initiated and which should not.

This type of approach has been adopted in the past in several research areas and contexts. See, for example, the expectancy-value class of attitude models in Fishbein (1961) and Rosenberg (1956), the composite criterion model proposed by Srinivasan and Shocker (1973), and investigated by Parker and Srinivasan (1976), and additive utility function examined in Keeney and Raiffa (1976), and Hwang and Yoon (1981). The Analytic

Hierarchy Process (Saaty, 1980) is another well known example of the use of linear composite indices for evaluating alternatives. In this setting, the parameters $\{v_{ij}\}_{ij}$ and $\{W_j\}$ are the normalized principal eigenvectors of appropriate ratio-scale matrices which represent pairwise comparisons of criteria importance and alternatives' relative standings with respect to each criterion.

In the usual setting, the criteria j are quantitative or numerical. Specifically, each alternative i is assigned a *cardinal* value v_{ij}. For example, the v_{ij} might be cost figures. Even when a criterion is inherently qualitative, such as the flavor of a product expressed on a ten point scale, the scale values v_{ij} are treated as if they were numerical in the same sense as cost data. (See, for example, the AHP model of Saaty (1980)). Korhonen and Wallenius (1990) examine the use of qualitative data in a linear decision model context. AHP techniques are used to estimate the linear coefficients of this model.

There are two major weaknesses with traditional approaches to MCDM. First, as indicated above, qualitative data in the form of rank positions is commonly treated as if it possessed quantitative meaning. Second, these frameworks do not have a convenient mechanism for handling simultaneously, a mix of qualitative and quantitative factors.

The models presented herein for dealing with MCDM structures is based on the DEA ideas of Charnes et al. (1978). Full details are provided in Cook and Kress (1991; 1994). In Section 7.2 we present brief descriptions of a number of case examples of problems involving multiple criteria. Section 7.3 develops the relevant model structures for tackling such problems. Section 7.4 extends such structures to the situation where some alternatives may be examined only in terms of proper subsets of the full set of criteria. Section 7.5 presents concluding remarks.

7.2. CASE EXAMPLES

7.2.1 Evaluating Capital Construction Projects

Many real world problems involve prioritizing a set of available alternatives, generally for the purpose of isolating a best or most desirable subset. A common example and possibly the most familiar one is that involving the ranking and selection of a set of fundable projects in a constrained budget situation. While any number of real world problem situations can be cited, we use two site specific project prioritization examples to illustrate the model presented later.

Prioritizing capital construction projects in a hydroelectric utility company

The annual capital construction program in a large hydroelectric company spans a broad range of initiatives – construction of new buildings, installation of power lines, upgrading of generating stations, and so on. These initiatives address many different needs and must be evaluated along several dimensions. The relevant dimensions include:

1. installation cost – this factor varies widely from project to project and may be a single-year or multiple-year value.
2. operating cost – this is an estimate of the ongoing cost of maintaining or operating the project or structure;
3. environmental impact – this factor is intended to capture the overall contribution to or detraction from the environment (air quality, ground water damage, and so on);
4. contribution to new energy sources and supply – some initiatives such as the installation of power lines can improve energy availability;
5. impact on existing or ongoing initiatives – certain projects may have an effect on existing programs, either in a positive or negative sense;
6. senior management support – this factor needs to be considered as it can influence the ongoing maintenance of the outcome from the project.

The first two criteria are clearly quantitative in the sense that a monetary figure, perhaps estimated, can be supplied for each initiative considered. The latter four criteria would, however, in most cases be considered as qualitative. A criterion for which precise quantitative (numerical) values can be obtained is called a *cardinal criterion.* A criterion which is qualitative in nature and according to which only a rank ordering of preferences can be obtained, is called an *ordinal criterion.* Specifically, in terms of a factor such as senior management support, one can only rank order or *categorize* the initiatives. For example, in the case of the hydroelectric company examined, five rank positions or categories are allowed for each ordinal criterion. In their terminology, these categories are designated

1. extremely important/valued;
2. very important/valued;
3. average importance/valued;
4. minor importance/valued;
5. not important/valued.

Therefore, each potential project is assigned a rank position from 1 to 5.

Having ranked the projects according to each criterion, the next step is to prioritize the criteria themselves. While in some environments the criteria may only be ordinally ranked (see Cook and Kress, 1991), in other cases numerical *criteria weights* are specified. The latter practice is common in situations where multiple criteria decision making is a recurring exercise. In a case such as the present one where prioritizing of a capital construction program is an annual exercise, it is very common for management-specified weights to be available. Furthermore, such weights are closely scrutinized and are re-evaluated on an ongoing basis as situations and priorities evolve.

What is also an important consideration in such decision making environments is the *clearness* or preciseness with which projects can be categorized in the case of an ordinal criterion. It may, for example, be easier to distinguish or discriminate between projects relative to environmental impact (criterion 3) than is the case for a criterion such as the impact on existing programs (criterion 5). This means that one would like to attach more relevance to the rank positioning in the case of the former criterion than in the latter. While existing composite index models do not directly account for criteria clearness, the methodology proposed herein contains such a facility.

Rehabilitation and system expansion decisions in highway network management

Transportation and highway departments everywhere are responsible for the management of the various highway networks under their jurisdictions. Highway capital expenditures fall into two general categories, namely rehabilitation of existing pavement and construction of new links (system expansion).

Many models exist which attempt to address the rehabilitation side of the highway management problem. These formal models have all attempted to view the problem of resource allocation as a single criteria problem, using an aggregate measure such as the PCI (Pavement Condition Index). Linear programming models, Markov models and various ranking procedures are some examples of the model structures used.

In regard to system expansion initiatives less formal attempts have been made. Although it has long been recognized that multiple criteria should be considered, final decisions on initiatives to be funded have generally been made in terms of some single factor such as 'capacity', deemed to be the *most important.*

In a recent study by the transportation department under consideration, a number of different factors or criteria were cited as being pertinent to decisions on priorities for funding. This realization has therefore raised the question as to how these various factors can be incorporated into a priority setting framework. The model described later has been tested on a large sample of data.

In the case of rehabilitation the set of criteria identified in the study are:

- PCI (pavement condition index);
- Present traffic;
- Predicted future traffic;
- Percent commercial vehicles;
- Accident level;
- Vehicle operating costs;
- Rider disruption.

In all but the last two criteria, numerical data is available. Since roads can only be ranked according to rider disruption, and since data on vehicle operating costs can be difficult to obtain, the latter two criteria are ordinal. For system expansion at least four major factors have been identified which should enter the priority setting exercise:

- Level of service (service is measured in six categories: A,B,C,D,E,F);
- Volume/capacity ratio;
- Accident prevention;
- Long run operating cost.

Here again some criteria are ordinal, while others are numerical.

In the case of both rehabilitation and system expansion there is general agreement regarding the relative importance of criteria. Specifically, it is believed that PCI is more important that the percent of commercial vehicles, which is more important than rider disruption, and so on. There is a further concern and belief that some criteria are 'clearer' than others. For example, it is easier to distinguish between the importance of the various rehabilitation sections on the basis of rider disruption than is the case for vehicle operating cost. The multidimensional nature of the problem is evident, as is the fact that both numerical and ordinal data are involved.

In the sections to follow we propose an extension to the usual composite index methodology. The main feature of this extension is that it is able to accommodate both ordinal and cardinal criteria. At the same time, it derives weights corresponding to the different rank positions or categories mentioned above, and addresses directly the issue of criteria clearness (or fuzziness). The ultimate outcome from the model is a logical means of deriving an aggregate or composite index for each alternative, hence a prioritization of the complete set of alternatives.

7.2.2 Selecting Automated Test Equipment at Northern Telecom

Northern Telecom (now called Nortel) had by the early 1990s established itself as a major player in the worldwide manufacture of digital switching equipment and related products that comprise the heart of telephone networks around the world. In 1990, discussion began between management and engineers in three divisions of Northern Telecom, involving the development of a new generation of circuit board testing equipment called Automated Test Equipment (ATE). New test equipment that was both efficient and flexible was needed to accommodate rapid changes in circuit design and customer demands for increasingly high levels of operating reliability.

In the past, test equipment had been developed in-house by one division, then bought and modified by other divisions for their needs. All three divisions produced different mixes of products and circuit boards. Maturation of the telecommunication industry towards common communication standards, however, has presented the opportunity for the organization to seek commonality of equipment between divisions, and to explore co-development with suppliers as well as off-the-shelf purchase of new equipment.

The emerging vision of the ATE is that of a system comprised of a sophisticated combination of hardware and software offering flexibility in the types and volumes of circuit packs that it could interface, and the conditions that could be simulated to test each board. It was felt that any specifications should be presented in a very loose fashion, and be of a qualitative rather than strictly quantitative nature. This would encourage individual prospective suppliers to exercise creativity in the deployment of their organization's unique capabilities. For example, the organization was interested in incorporating artificial intelligence software into the design. To operationalize this into a specification, a description was provided as to what this software would do, where it would be used and the user interface. Individual vendors would respond with what they believed they could offer in the way of making this vision a reality. Thus, two vendors could have two different designs but with equal merit. By allowing suppliers to promote their own unique technical capabilities, it was believed that the overall design would be improved in the long run. Northern Telecom, therefore, realized that they would have to be flexible in the definition and evaluation of design criteria.

In evaluating vendors, experience in the telephone industry and familiarity with ATE technology were highly valued. These two general criteria were used to informally reduce a field of many potential suppliers down to a manageable set, each member of which would be asked to submit proposals for participation in a co- development relationship. All three divisions had preferred vendors that they felt should be considered, but all agreed that invited vendors should have some experience in the industry to facilitate communication between technical personnel. Northern Telecom believed it could bring extensive skills and experience to any relationship, but required that all potential vendors have familiarity with the building of integrated test systems.

Six alternatives were considered; three involved in-house development, and three with other companies. Each organization submitted detailed information about their approach to each specification. In addition, they were asked to cost their proposals as best they could, given the preliminary nature of the discussions. Each of the three divisions involved in the ATE team then evaluated the information received. To establish a common value system, the evaluation team had developed a decision matrix which accommodated most of the issues of interest to all three of the divisions. A total of 40 criteria covered the spectrum of issues deemed important by the three divisions.

The decision matrix in Table 7-1 spells out the "rating" of each supplier along each criterion. A 5-point Likert scale was used to specify the level of importance in each case.

The methodology available at that time for combining outcomes across the criteria involved the supplying of "weights" as shown.

To derive an overall rating for each of the six alternatives (vendors), a weighted total score was computed. For alternative No. 1, for example, the rating (on a one to five scale) for criterion No. 1 is 4, for criterion No. 2 is 5, etc. The criteria weights are: criterion No. 1 - 8; criterion No. 2 - 10; $\cdots$. Thus, the overall score for alternative No. 1 is $8 \times 4 + 10 \times 5 + \cdots = 869$. The overall scores resulting from the data of Table 7-1 are

Vendor #	Rating
1	869
2	876
3	1213
4	1360
5	1103
6	1059

Therefore, the relative ranking of the vendors that one achieves by applying this weighting method to the data of Table 7-1 are $No.\,4 > No.\,3 > No.\,5 > No.\,6 > No.\,2 > No.\,1$.

Table 7-1. Evaluation Matrix

Criteria #	Evaluation Weights	Supplier 1	2	3	4	5	6
1.	8	4	5	5	5	3	1
2.	10	5	1	5	3	5	4
3.	8	2	1	1	5	2	4
4.	8	5	5	5	3	5	4
5.	7	2	2	1	2	5	3
6.	10	2	2	2	5	4	4
7.	10	2	3	3	4	4	4
8.	7	2	3	4	5	3	3
9.	10	2	2	4	4	2	2
10.	10	5	5	5	1	1	1
11.	10	1	1	5	5	4	3
12.	10	1	1	4	4	4	1
13.	8	1	1	2	5	1	3
14.	6	1	1	2	3	1	3
15.	10	5	5	5	5	5	3
16.	8	3	2	3	4	3	2
17.	8	1	1	2	5	2	5
18.	8	1	1	2	5	2	5
19.	10	5	5	5	5	5	4
20.	8	3	2	3	5	3	4
21.	6	1	1	1	5	2	3
22.	6	1	1	1	2	1	3
23.	7	1	1	1	1	1	1
24	9	2	5	5	5	5	3
25.	9	1	1	1	1	1	1
26.	7	2	2	4	4	4	4
27.	10	5	5	5	5	5	5
28.	7	3	3	5	5	5	5
29.	9	1	1	1	1	1	1
30.	9	1	1	1	1	1	1
31.	9	5	5	5	5	5	5
32.	10	4	5	5	5	2	2
33.	9	2	3	4	4	4	4
34.	8	1	1	4	4	4	3
35.	10	2	2	3	3	2	3
36.	10	2	2	5	5	4	4
37.	10	3	3	5	5	3	3
38.	10	1	1	5	5	5	5
39.	5	1	1	1	4	2	4
40.	10	5	5	5	4	4	1

Note: The specific nature of the criteria are not specified for reasons of confidentiality.

This "supplied weighting" approach is one that is commonly adopted in such multiple criteria problems. While this approach is simple to understand, it possesses at least two major weaknesses. First, one is forced to specify explicitly the numerical sizes of weights, using some arbitrarily chosen scale. This exercise is very much at the whim of the decision maker(s), even when it is based on the very best advice and information from the relevant players at the time. In the case at hand, the weights were arrived at through a process which, although democratic and consensus seeking in nature, resulted in widely varying opinions as to what the relative sizes of the weights should be. In this regard, the extent to which the weights reflect a true "consensus" may be less than satisfactory.

The second, and even more disturbing aspect of the methodology used to rank the vendors, has to do with the treatment of the 5 point ordinal scale on which vendors are ranked as if it were a cardinal (interval) scale. Specifically, the overall score for a vendor was taken as a weighted sum of the rank positions which that vendor achieved. This, however, is generally not the intention of these rank positions. For example, in stating that vendor A is preferred to vendor B, and B is preferred to C (according to some criterion), and assigning them rank positions 4, 3 and 2 respectively, should not carry with it the connotation that "A is twice as importance as C" (i.e. having a rating of 4 versus a rating of 2). The numbers 4, 3 and 2 should be interpreted as *rank positions only,* and not as absolute worths of A,B and C according to the criterion in question.

In the section to follow, we present an alternative approach for achieving a prioritization of the vendors, where only ordinal or ranking data is available. The approach makes use of the ordinal scales only in *that* sense, as distinct from the above described procedure which explicitly uses the numbered rating data. Furthermore, the approach does not require that the decision maker explicitly specify criteria weights, but rather that only the relative importance of the criteria be given.

7.2.3 Country Risk Evaluation

Country risk evaluation is an important component of the investment and capital budgeting decisions of international investors. The increased internationalization of investment in recent decades has raised the exposure of investors to the risks associated with events in many different countries. Consequently, substantial resources are now being devoted to country risk analysis by international investors who realize the importance of identifying, evaluating, and managing the uniquely international risks they face. For many international investors, profits and opportunities for growth have increasingly come to depend on how effectively they cope with international uncertainties.

In response to increased internationalization of investment, formal country risk evaluation is becoming "firmly established as one of the essential international business functions" (Ting, 1988). For example, country risk assessment units have been established by most large financial institutions and by more than one-third of the companies contacted in a recent survey (Kobrin, 1986). However, typically the unit's approach is neither formal nor analytical, and may often involve no more than a checklist or the outline of a country study. Few companies appear to have developed a systematic method of assessing the uniquely international risks facing their various projects.

While assessing and quantifying the impact of the country risks facing a MNCs (Multinational Corporations) operations presents a formidable problem, it is an essential component of a MNCs capital budgeting decisions. For example, to determine either the appropriate risk-adjusted discount rate to use or how to correctly adjust expected cash flows, it is necessary to evaluate and quantify the relevant country risk factors, and to monitor their changes over time (see Ang and Lai (1990), and Goddard (1990)). Then capital budgeting decisions can directly utilize comparable evaluations of country risk, and minimize the subjective component of the risk assessment process. Further, a formal, analytical approach facilitates testing to determine which approach and what factors have been the most successful in correctly assessing the importance of the various risk criteria (see Blask, 1978; Ting, 1988).

There are two broad categories of risk faced by international investors: "macrosociopolitical" risks and "micro" risks. The potential impact of both macro- and micro-risks vary among specific projects, firms and industries. The extent to which a MNC's projects are vulnerable to these risks depends as much, if not more, on the specific nature of the project as on the condition of the host country. For this reason it is clearly inappropriate to adopt a single country risk measure for all MNCs, or for all projects of a particular MNC, in a specific country (this point is the central theme of Ting (1988) and, as he notes, the most common criticism of country risk-rating services).

The value of a project's "assets in place" are largely independent of the MNC's future investment strategy and of the project's future ownership structure. The value of a project's "real options" however follows from the MNC's discretionary exercise of options to exploit positive NPV investment opportunities. These real options are largely firm specific (requiring, for example, the MNC's idiosyncratic technical skills or manufacturing processes) and hence their value is highly dependent on both the MNC's future investment strategy and on the project's future ownership structure. Therefore, Phillips-Patrick (1989, 1990) argues that a project whose value consists primarily of "assets in place" is subject to greater country risk than a project whose value consists mainly of "real options."

Currently, the most popular quantitative approach to country risk evaluation applies fixed (often equal) weights to the risk variables or criteria employed. The approach developed in this chapter allows investors to rank the risk criteria themselves, according to both their importance and their relative clarity. This approach recognizes that the rankings of the criteria will, in general, vary from one investor to another, reflecting the heterogeneous nature of their projects. Since the investors' projects possess heterogeneous sensitivities to the various types of risk, they will, in general, rank the various criteria (risks) differently. This allows each investor to determine a different optimal set of criteria weights and hence, to obtain a different rating (and hence, ranking) of the riskiness of the countries evaluated, dependent upon the specific nature of their investment project.

Cook and Hebner (1993), examine 14 criteria for evaluating risk. Data obtained from the Japanese Bond Research Institute was used as the basis for ranking 100 countries in terms of investment risk. The specific criteria employed were:

i) F_i^1 - social stability rating for country i.
ii) F_i^2 - political stability rating for country i.
iii) F_i^3 - consistency of policies rating for country i.
iv) F_i^4 - industrialization rating for country i.
v) F_i^5 - economic problems rating.
vi) F_i^6 - fiscal policy rating.
vii) F_i^7 - monetary policy rating.
viii) F_i^8 - growth potential.
ix) F_i^9 - susceptibility to war.
x) F_i^{10} - international standing.
xi) F_i^{11} - balance of payments.
xii) F_i^{12} - debt servicing capacity.
xiii) F_i^{13} - foreign investment policy.
xiv) F_i^{14} - foreign exchange policy.

In the section to follow a model structure is presented that can be utilized to tackle practical MCDM problems of the type discussed above. For further details on this model, see Cook and Kress (1996).

7.3. THE MODEL

Suppose N alternatives (for example, projects) are to be evaluated relative to a set of K_1 ordinal and K_2 cardinal criteria. Denote these two sets of criteria as ORD and CARD, respectively. Thus,

$$K_1 = |\text{ORD}| \text{ and } K_2 = |\text{CARD}|.$$

For the k-th cardinal criterion, let $a_k(i)$ denote the value or worth associated with alternative i. As in the AHP case (Saaty (1980)), there is no loss of generality in assuming that the $\{a_k(i)\}_{i=1}^{N}$ are normalized, that is $\sum_{i=1}^{N} a_k(i) = 1$. Let W_k denote the supplied weight or importance associated with criterion k. Again without loss of generality, we assume

$$\sum_{k=1}^{K_1+K_2} W_k = 1. \qquad (7.1)$$

Now, in the case that only cardinal criteria are present, the *composite index* or aggregate rating corresponding to alternative i is given by $\sum_{k \in \text{CARD}} W_k a_k(i)$. Since the $\{a_k(i)\}_i$ are normalized, the term $W_k a_k(i)$ is the proportion of W_k credited to the i-th alternative.

For ordinal criteria k, there are no specified $a_k(i)$- values. What is supplied is the rank position of the i-th alternative on the k-th criterion. In that regard, define

$$d_{k\ell}(i) = \begin{cases} 1 \text{ if alternative } i \text{ is ranked in } \ell\text{-th position} \\ \text{on criterion } k, \\ 0 \text{ otherwise.} \end{cases}$$

To see how a composite index should be defined when both ordinal and cardinal criteria are present, we argue as follows:

$$W_k' = \theta W_k, \qquad (7.2)$$

where θ is a scaling parameter to be determined by the model below. Thus, the W_k' are simply scaled versions of the W_k. For ordinal criterion k, let $w_{k\ell}$ denote the value or worth associated with rank position ℓ on criterion k (see Cook and Kress, 1991). For consistency, the $w_{k\ell}$ -values should take on the same role for ordinal criteria as the $a_k(i)$ assume for cardinal criteria. More to the point, if the rank position occupied by the i-th alternative relative to criterion k is denoted by ℓ_i, then $w_{k\ell}$ is analogous to $a_k(i)$. Note that

$$w_{k\ell_i} = \sum_{\ell=1}^{L} d_{k\ell}(i) w_{k\ell}.$$

Clearly then, to carry the composite index idea over to the ordinal setting it is reasonable to impose the constraint

$$\sum_{\ell=1}^{L} \sum_{i=1}^{N} d_{k\ell}(i) w_{k\ell} = \sum_{\ell=1}^{L} M_{k\ell} w_{k\ell} = 1, \quad k \in \text{ORD}, \qquad (7.3)$$

since

$$\sum_{i=1}^{N} a_k(i) = 1 \quad \text{for any } k \in \text{CARD}.$$

Here,

$$M_{k\ell} = \sum_{i=1}^{N} d_{k\ell}(i).$$

With these ideas in place, we extend the definition of a composite index to the mixed ordinal/cardinal case.

Definition 7.1: In the presence of both ordinal and cardinal criteria, the mixed criteria composite index for an alternative is given by

$$R_i = \sum_{k\in\text{ORD}} \sum_{\ell=1}^{L} W_k' d_{k\ell}(i) w_{k\ell} + \sum_{k\in\text{CARD}} \text{W}_k' a_k(i). \tag{7.4}$$

It is noted than an ordinal criterion contributes

$$W_k' \sum_{\ell=1}^{L} d_{k\ell}(i) w_{k\ell} = \theta W_k \sum_{\ell=1}^{L} d_{k\ell}(i) w_{k\ell}$$

to the index R_i. A cardinal criterion contributes $\theta W_k a_k(i)$.

To facilitate the discussion below, we introduce a change of variables

$$x_{k\ell} = \theta w_{k\ell}, \text{ for } k \in \text{ORD}. \tag{7.5}$$

The composite index is then rewritten as

$$R_i = \sum_{k\in\text{ORD}} \sum_{\ell=1}^{\text{L}} \text{W}_k d_{k\ell}(i) x_{k\ell} + \theta \sum_{k\in\text{CARD}} \text{W}_k a_k(i) \tag{7.4'}$$

where

$$\sum_{\ell=1}^{L} M_{k\ell} x_{k\ell} = \theta, \tag{7.6}$$

due to (7.3).

The issue now arises as to how to determine an appropriate set of $x_{k\ell}$ values as well as an appropriate scaling parameter θ. First we discuss constraints that the $x_{k\ell}$ should satisfy, and describe how the criteria fuzziness concept can be incorporated into the $x_{k\ell}$ derivation process.

Rank position discrimination

For any criterion k the importance to be associated with the ℓ-th rank position should exceed that of the $(\ell+1)$ st. Therefore, it should be true that $x_{k\ell} > x_{k\ell+1}$. More particularly, define $G_{k\ell}$ to be a positive function, and impose the constraints $x_{k\ell} - x_{k\ell+1} \geq G_{k\ell}$ for all k, ℓ. The $G_{k\ell}$ should reflect two phenomena. First, the *relative* worths $x_{k\ell}$ of the successive rank positions $(\ell = 1, 2, \ldots, L)$ may not be equally spaced. That is, it may be desirable to have larger gaps $(x_{k\ell} - x_{k\ell+1})$ between some pairs of rank positions ℓ and $\ell+1$ than between other pairs. It may, for example, be desirable to distinguish more clearly between the first and second rank positions than between the ninth and tenth positions. Second, in addition to accounting for the relative *gaps* between rank positions, the $G_{k\ell}$ should embody a lower bound on the *absolute gaps* as well. To accommodate both aspects, represent $G_{k\ell}$ in a product form $G_{k\ell} = g_\ell u_k$. The g_ℓ are supplied parameters which are to capture the lower bounds on the relative gaps $x_{k\ell} - x_{k\ell+1}$. If, for example, we wished to distinguish less and less between rank positions as ℓ increases, then the g_ℓ would be chosen as a decreasing sequence. The u_k are decision variables to be determined by the model (see below). As will be discussed below, these variables will be used to reflect criteria clearness. The product $g_\ell u_k$ provides, then, the minimum absolute discrimination between consecutive rank positions for an ordinal criterion k.

Criteria clearness

The u_k provide a convenient means of capturing criteria clearness. In particular, if criterion k_1 is clearer than criterion k_2, we impose the constraint $u_{k_1} > u_{k_2}$ (or $u_{k_1} - u_{k_2} - z \geq 0$, where z is a small positive scalar). Thus, smaller u_k values correspond to fuzzier criteria, meaning that for any rank position $\ell, G_{k_1\ell}$ is larger than $G_{k_2\ell}$. Hence, the minimum amount of discrimination between consecutive rank positions ℓ and $\ell+1$ is greater for clear criteria than for fuzzy criteria.

Having established certain restrictions to which the $x_{k\ell}$ should adhere, a reasonable approach for deriving a composite index for an alternative i is to find a set of $x_{k\ell}$ (and θ) that meet these restrictions, but that also show alternative i in the most favorable light. This philosophy was introduced by Charnes, Cooper and Rhodes (1978) in the context of developing efficiency measures for a set of decision making units (DMUs). In the present setting, the alternatives or projects would represent the decision making units. Using this argument, an appropriate mixed-criteria composite index can then be derived for an alternative i_o by solving the linear programming problem:

$$R^*_{i_o}(z) = \max \sum_{k \in \mathrm{ORD}} \sum_{\ell=1}^{\mathrm{L}} \mathrm{W}_k d_{k\ell}(i) x_{k1} + \theta \sum_{k \in CARD} W_k a_k(i) \quad (7.7)$$

subject to

$$\sum_{k \in \mathrm{ORD}} \sum_{\ell=1}^{\mathrm{L}} \mathrm{W}_k d_{k\ell}(i) x_{k\ell} + \theta \sum_{k \in CARD} W_k a_k(i) \leq 1, \; i = 1, \ldots, N, \quad (7.8)$$

$$x_{k\ell} - x_{k\ell+1} - g_\ell u_k \geq 0, \;\; k \in \mathrm{ORD}, \;\; \ell = 1, \ldots, L-1, \quad (7.9)$$

$$x_{KL} - g_L u_k \geq 0, \quad k \in \text{ORD}, \tag{7.10}$$

$$\sum_{\ell=1}^{L} \text{M}_{k\ell} x_{k\ell} - \theta = 0, \quad k \in \text{ORD}, \tag{7.11}$$

$$u_{\langle j \rangle} - u_{\langle j+1 \rangle} \geq z, \quad j \in \text{ORD}, \tag{7.12}$$

$$u_{\langle K_1 \rangle} > z,$$

$$x_{k\ell}, \theta, u_k \geq 0. \tag{7.13}$$

We, therefore, find variables $x_{k\ell}, \theta$ and u_k which yield the highest possible index R_{i_o} for the alternative i_o in question, subject to these variables abiding by a set of natural restrictions. This problem is solved for each one of the N alternatives. Constraints (7.9) provide for the lower limit restraints discussed earlier. Constraints (7.11) incorporate the fuzziness facility. In this notation, $u_{\langle j \rangle} = u_k$ if criterion k is ranked j-th in terms of clearness. Constraing (7.10) is simply (7.6) rewritten. In order to bound the problem, constraints (7.8) impose an upper limit on each composite index. Clearly, any bound could be imposed here, but since normalization to unity was used on the raw data ($a_k(i)$ and W_k), it is natural to use an upper limit of 1.

We have used the notation $R_{i_o}(z)$ to denote the fact that the solution obtained will generally depend on the value of z chosen. Clearly, as is true in the data envelopment analysis context, a different set of $\{x_{k\ell}, \theta, u_k\}$ can arise for each alternative i. While this inherent property of flexible weighting parameters is theoretically sound, it has raised some objections from practitioners who have found difficulty in accepting the principle that one DMU is viewed in a different perspective than other DMUs. In the engineering context, for example, productivity measurement is based on finding a common set of standards or multipliers, hence the idea of flexible weights does not arise.

Thus, while the idea of weight flexibility is appealing in that it helps to show each alternative in its best light, practical considerations may dictate that a common set of weights be determined. In the general DEA context, the idea of a common set of weights has been examined by Roll et al. (1991).

In the following section it is shown that the n-problem structure of model (7.7)-(7.12) can be replaced by a single-problem structure. This structure which yields a single set of weights is shown to be equivalent, in a certain sense, to the original n-problem model.

7.3.1 The Modified Model

The model of Section 7.3 possesses a number of important properties as spelled out by the theorems below. In the presence of these properties it will be shown that the basic model can be replaced by a simple modified version to affect a ranking of the alternatives.

Theorem 7.1. The optimal rating scores $R_i^*(z)$ of (7.7) are monotonic non-increasing in z.

Proof: This follows from the fact that increasing values of z further constrain the linear programming model. Q.E.D.

Theorem 7.2. If z_i^* is the maximum value of z for which $R_i(z)=1$, and if $z_{\max} = \max_i \{z_i^*\}$, then problem (7.7)-(7.12) is feasible if and only if $z \le z_{\max}$.

Proof: Suppose there exists a value $\hat{z} > z_{\max}$ for which (7.7)-(7.12) is feasible, and that $\hat{x}_{k\ell}, \hat{\theta}, u_k$ is a feasible solution for that $\hat{z}$. By definition of $z_{\max}, R_i < 1$ for all i. Let $\hat{\varepsilon}$ be a small parameter, and let $\{\hat{\varepsilon}_{k\ell}\}$ be a non negative solution to the set of linear systems of inequalities.

$$\sum_{\ell} \mathrm{M}_{k\ell}\varepsilon_{k\ell} = \hat{\varepsilon}, \tag{7.14a}$$

$$\varepsilon_{k1} \ge \varepsilon_{k2} \ge \ldots \ge \varepsilon_{k\ell} \ge 0. \tag{7.14b}$$

It is easily seen that a solution for (7.14a)-(7.14b) always exists. Specifically, define:

$$\tilde{x}_{k\ell} = \hat{x}_{k\ell} + \hat{\varepsilon}_{k\ell},$$
$$\tilde{\theta} = \hat{\theta} + \hat{\varepsilon},$$

where $\hat{\varepsilon}$ is small enough such that $R_i \le 1$, for all i. Clearly, $\{\tilde{x}_{k\ell}\}, \tilde{\theta}$ also constitutes a feasible solution for (7.7)-(7.12) since

(a) $R_i \le 1$, for all i, by definition.

(b) $\tilde{x}_{k\ell} - \tilde{x}_{k\ell+1} = \hat{x}_{k\ell} + \hat{\varepsilon}_{k\ell} - \hat{x}_{k\ell+1} - \hat{\varepsilon}_{k\ell+1} \ge \hat{x}_{k\ell} - \hat{x}_{k\ell+1},$ since $\hat{\varepsilon}_{k\ell} \ge \hat{\varepsilon}_{k\ell+1}$.

(c) $\sum \mathrm{M}_{k\ell}\tilde{x}_{k\ell} - \tilde{\theta} = \sum \mathrm{M}_{k\ell}\hat{x}_{k\ell} + \sum \mathrm{M}_{k1}\hat{\varepsilon}_{k\ell} - \hat{\theta} - \hat{\varepsilon} = \sum \mathrm{M}_{k\ell}\hat{x}_{k\ell} - \hat{\theta} = 0.$

By choosing $\hat{\varepsilon}$ large enough to drive at least one of the R_i up to 1, we have a feasible solution with at least one of the $R_{i_o} = 1$, that is $R_{i_o}(\hat{z}) = 1$, in contradiction to the maximality of $z_{\max}$. Q.E.D.

Theorem 7.3. If $R_i^*(z^*) = 1$, for all $z \le z^*$.

Proof: This follows from Theorem 7.1. Q.E.D.

Theorem 7.4. At optimality, at least one of the constraints of (7.8) is binding.
Proof: The value of z is restricted by the constraints in (7.8) (the problem Max z, subject to (7.9)-(7.12) is unbounded). If all hold at strict inequality (LHS strictly less than 1), then it means that z can be increased, in contradiction to its maximality. Q.E.D.

From the above theorems and observations, it is clear that different values of the parameter z can lead to different solutions. The larger the value of z chosen, the greater become the gaps between the importance values $x_{k\ell}$ assigned to consecutive rank positions on ordinal criteria. As indicated earlier, z is a minimum measure of discrimination between such rank positions. It is also true that the larger the value chosen for z, the more constrained the feasible region becomes, and the fewer will be the number of alternatives achieving a rating of 1. (See Theorem 7.1).

There are two compelling arguments for using the largest possible value of z(i.e., $z_{\max}$) in deriving ratings for the alternatives (that is, in solving (7.7)-(7.12)). First, there is the issue of discrimination as raised in the preceeding discussion. Since the intention of the model is to assign worths or values to the chosen rank positions $\ell = 1, 1, \ldots, L$, it is reasonable to differentiate between these positions to the greatest extent possible. The value $z_{\max}$ accomplishes this goal. Second, from Theorem 7.3, any alternative that achieves the rating status of 1 for $z_{\max}$ will retain this status for all smaller z-values. Such alternatives are in this sense, then, truly deserving of a *first-place* status as compared to alternatives that were in first place for smaller values of z but lost this status for larger values. Hence, those alternatives i with $R_i(z_{\max}) = 1$ are the real *winners* in the ultimate ranking scheme.

We, therefore, propose a modified version of model (7.7) through (7.12) which finds the maximum value of the parameter z for which a feasible solution exists. Specifically, we solve the problem

$$\begin{aligned} &\max \quad z \\ &\text{subject to constraints (15.8)-(15.12),} \end{aligned} \tag{7.15}$$

where z is treated as a variable rather than as constant in constraints (7.11). Again, we point out that in solving such a problem we are finding a solution (a set of variables $(x_{k\ell}, \theta, u_k)$) which possesses the highest degree of discrimination between criteria on the basis of clearness, hence the greatest discrimination between rank positions ℓ.

It has been found that in many cases there are alternate optima to this problem. While there is no clear indication that any one solution to such a problem is preferred over any other, it can be argued that a solution which makes as many of the R_i as large as possible would be highly desirable. Consequently, the objective function of (7.2) can be replaced by

$$\max \ Mz + \sum_{i=1}^{N} R_i, \tag{7.16}$$

where M is a large positive number.

In this manner, the largest degree of discrimination between criteria and rank positions is achieved (z is maximized) and the total of the composite indices is maximized at the same time. It must, of course, be pointed out that because of the existence of alternate optima, certain alternatives may be ranked in first place using one solution, while others may possibly occupy first place if a different solution is used. (7.16) does provide an opportunity, however, to find a solution with as many first place alternatives as possible.

To illustrate the modified model, consider the following numerical example based on the capital construction problem described in Section 7.2.

Example: The following criteria weights approximately reflect the priorities set by an electrical utility company.

(1) Initial cost = 0.12
(2) Operating cost = 0.08.
(3) Environmental impact = 0.20
(4) Contribution to new sources = 0.15
(5) Impact on existing activities = 0.10
(6) Management support = 0.35.

The ratings of 10 proposed projects along each of the six factors are given in Table 7-2.

Table 7-2. Ranking of alternatives

Alternative	Criteria					
	1	2	3	4	5	6
1	10	1.2	1	1	3	4
2	5	1.7	1	3	4	4
3	2.3	2.3	4	4	2	3
4	17	2.6	5	1	1	1
5	12	0.8	2	5	2	2
6	32	0.3	3	4	3	5
7	12	3.5	4	5	5	5
8	19	6.2	2	2	3	2
9	6.1	2.1	3	1	5	1
10	5	2.5	4	1	2	4

It is noted that the numerical values for the first two criteria are in fact figures which have been obtained by subtracting the actual costs from a value larger than the maximum cost in each criterion.

When the modified model of this section was applied, the resulting composite indices and rankings obtained were as shown in Table 7-3. In a second run of the model, the weights on criteria 1 and 3 are interchanged. The results are displayed in Table 7-4.

Table 7-3. Composite indices and rankings of alternatives

Alternative	Composite indices (%)	Rankings
1	79.9	4
2	71.2	8
3	61.9	9
4	100.0	1
5	77.0	5
6	75.8	6
7	50.8	10
8	100.0	1
9	92.1	3
10	72.8	7

Table 7-4. Revised indices and rankings

Alternative	Composite indices (%)	Rankings
1	73.1	5
2	61.0	8
3	54.8	9
4	100.0	1
5	73.0	6
6	87.2	3
7	51.8	10
8	100.0	1
9	76.8	4
10	68.7	7

From a comparison of the two tables we can see that a number of the alternatives are still ranked at the same level. Alternatives 4 and 8, for example, retain their 100% status. This is reasonable since in the case of alternative 4, say, less importance is now attached to its fifth place standing on the third criterion. Thus, the 100% rating is even more justified after the change in weights than before. Alternative No. 6 went from a sixth place standing to a third place standing due primarily to a heavier weight being shifted to criterion No. 1 where alternative 6 is a top performer.

A third analysis was performed in which criterion No. 4 was removed from the problem and its weight equally distributed over the remaining five factors. The resulting output is shown in Table 7-5.

Note that a dramatic shift in the rank position of project No. 9 has occurred (it went from a ranking of 3 in Table 7-3 to a ranking of 9 in Table 7-5). This can be partially explained by the fact that a criterion on which this project was ranked in first place has been removed. Hence, the strength of this positioning has been lost.

In the above analyses, model (7.15) was used since the desire was to prioritize all projects in term of the same optimal solution. It is worth pointing out that if model (7.7)-(7.12) is applied somewhat different results arise. Applying the original criteria weights, the comparable results to those of Table 7-3 are shown in Table 7-6.

Table 7-5. Results with fourth criterion removed

Alternative	Composite indices (%)	Rankings
1	84.1	5
2	69.0	8
3	71.2	9
4	100.0	1
5	84.4	4
6	95.5	6
7	53.2	10
8	100.0	1
9	64.0	1
10	83.9	7

Table 7-6. Revised Composite Indices

Alternative	Composite indices (%)	Rankings
1	82.5	5
2	71.2	8
3	67.7	9
4	100.0	1
5	84.3	4
6	78.2	6
7	60.1	10
8	100.0	1
9	100.0	1
10	74.1	7

Since these indices arise as a result of finding a best set of weights for each alternative (using the optimal z from (7.16)) each rating will be at least as large as before. Clearly, since each alternative now has its rating maximized, the sum of ratings will 'truly' be maximized as opposed to attempting to maximize this sum in (7.16). Notice that the relative positioning of alternatives is roughly the same (but not exactly) as was the case in Table 7-3. Note also that in addition to the two alternatives with a previous rating of 100% retaining that rating, a third alternative has moved from 95.5% to 100%.

In summary, it must be said that while one can get a somewhat different picture of the *relative positioning* of the alternatives from the two models, the many examples run have shown that the differences are minor. The advantage that (7.16) has over (7.7)-(7.12) is that a single set of 'best' weights is found, hence having more managerial appeal.

7.3.2 Implementation Issues

In the two site-specific situations examined, a number of implementation issues deserve mention.

Prioritization versus resource allocation

This tool is designed as a means of rank ordering a set of alternatives (for example, projects). In a limited budget framework, the projects to be 'funded' are then selected by starting with the top ranked alternative and moving down the list until available resources are exhausted. The assumption is that each project is in a 'go – no go' situation; i.e., either it is funded 100% or not at all.

While in the capital construction project situation under consideration this is generally the case, there are settings where partial funding is an option. The issue as to how to use the ratings R_i to decide what portion of a project to fund is, however, an open question. One suggestion in the case of AHP (Saaty, 1980) is that the level of funding should be proportional to the weighted rating the project receives. It is generally the practitioners' view, however, that such an approach makes no sense in that the R_i are really only relative measures. Thus, the problem of resource allocation, where projects can be partially funded, seems to be an unresolved issue.

Non-comparable subsets of projects

A major problem arising in the highway project ranking problem is one involving 'non-comparable' projects. Specifically, rehabilitation projects must be evaluated in terms of one set of criteria, while system expansion projects are to be rated relative to a different set of factors. Admittedly, some criteria such as job creation are common to both sets of projects. However, the cost of any project in the latter category generally far exceeds that for alternatives in the former category.

Having recognized this non-comparability issue, the organization is now attempting to gain a better understanding as to the appropriate criteria to use to bring the two categories of projects into a common light.

Rank reversal

The model, like other tools such as AHP, *does* exhibit the rank reversal phenomenon in certain situations. Recall that this means that the relative ranking of two alternatives can be reversed if a third alternative is removed from the set being evaluated. In a large number of examples examined, however, rank reversal occurred only a small percentage of the time and in situations where pairs of alternatives are ranked very closely (e.g., $R_1 = 0.79$ and $R_2 = 0.78$). In the two site-specific applications under study no serious rank reversal situation occurred.

7.4. EVALUATION RELATIVE TO PARTIAL CRITERIA

In the previous sections it was assumed that any given alternative i could be evaluated (assigned a rank position) in terms of each member k of the full set of criteria K. In many decision environments, however, this requirement is not pertinent. Consider, for example, the case where in ranking projects in an electric utility company, one may be considering alternatives such as construction of power lines, additions and modifications to nuclear reactors, upgrades to buildings, maintenance of office facilities, and so on. In such a varied set of alternatives, criteria such as "impact on environment," or, "contribution to technological advancement" may apply to some options (e.g., reactor construction), but may be entirely inapplicable to others such as building maintenance. In a completely different setting, consider one of the principal application areas of data envelopment analysis, namely the evaluation of productivity of a set of bank branches. See, for example, Sherman and Gold (1985) and Oral and Yolalan (1990). The traditional settings examined to date and cited in the literature, view banks at a given point in time and assume each branch can be evaluated in terms of the same criteria (inputs and outputs). If we want to compare, however, the new full service type of banking environment to the traditional branches, problems arise. The new style banks now offer services such as life and property insurance policies, mutual fund investment options, and so on, that are not available in some (conventional) branches. The comparison of old and new as a single set will then need to consider the partial criteria issue.

The problems associated with comparing a set of alternatives (projects, bank branches, etc.) when some criteria are relevant to certain members of the set but not to others, revolve around the interpretation of missing data and how to account for it. One ad hoc approach to this has been to generate synthetic data by using, for example, an average value for a criterion, where the average is over those alternatives for which that criterion is relevant. In the case of the bank branches, for instance, this would mean looking at the average of insurance sales for the new style branches, and then crediting each of the old style branches with that average value. In assessing projects, one option clearly is to fully penalize an alternative for "failing to perform" on a given dimension. Being fully penalized may mean being credited with the worst possible rank position on the given criterion, or being assigned no rank at all. This latter is the basis for the aggressive model to follow. On the other hand, if one argues that an alternative should not be penalized for not being eligible to be ranked on a given criterion, then a more benevolent action should be taken.

We now consider the general case in which an alternative i can be evaluated in terms of only a subset $K_i \subseteq K$ of the criteria. The manner in which the set of N alternatives is to be evaluated in this *partial criteria* case depends upon the assumptions one makes regarding *fair comparison.* We present three approaches to the evaluation:

Aggressive evaluation

One point of view regarding evaluation of the N alternatives is to adopt the original full criteria model ((7.7)-(7.13)), and replace the term $\sum_{k=1}^{K}\sum_{l=1}^{L} d_{kl}(i)w_{kl}$ by $\sum_{k \in K_i}\sum_{l=1}^{L} d_{kl}(i)w_{kl}$. In this case when a criterion k_o is not part of the pertinent set K_{i_o} for alternative i_o, a credit of 0 is given. That is $d_{k_o l}(i_o) = 0$ for all l. This approach subscribes to the concept that part of any alternative's worth (e.g., the worth of a project to an organization) is the benefit w_{kl} derived from each criterion. Hence, the fact that the project cannot compete in terms of a particular criterion k only serves to put that project at a disadvantage vis-à-vis other projects which *do* obtain a rank position on k. Thus, projects must compete *aggressively* (or at least are evaluated aggressively) with no compensation for failure to achieve a standing relative to certain criteria.

Clearly, this approach rewards those alternatives for which the cardinality $|K_i|$ of K_i is large, and penalizes those for which the cardinality is small.

While the approach has the advantage of treating all alternatives on an equal footing, it could be judged as being unfairly harsh in situations where criteria are simply inapplicable. In a situation, for example, where environmental impact is one of the factors used for evaluation, the $1 \rightarrow L$ scale may, in some circumstances, be interpreted as "good" to "bad." Thus, a rating of $l = 1$ means that an alternative has a very positive effect vis-à-vis environmental benefits, while $l = L$ may imply a very negative impact. An alternative (e.g., building maintenance) which is *neutral* should, if given a rank at all, be rated somewhere in the middle of the scale. Hence, the manner in which scales are defined can influence the applicability of the standard model in the partial criteria case.

Average performance evaluation

To avoid the potential problems created by cardinality differences among the sets K_i, as cited in the previous model, an approach which utilizes an *average performance* per pertinent criterion can be adopted. Specifically, we replace $\sum_{k=1}^{K}\sum_{l=1}^{L} d_{kl}(i)w_{kl}$ in (7.7) and (7.8) by $\sum_{k\in K_i}\sum_{l=1}^{L} d_{kl}(i)w_{kl} / |K_i|$. In a certain sense, this model is a natural extension of (7.7)-(7.13). That is, if in the full criteria case (i.e., $|K_i| = N$ for all i) we replace $\sum_{k=1}^{K}\sum_{l=1}^{L} d_{kl}(i)w_{kl}$ by $\sum_{k=1}^{K}\sum_{l=1}^{L} d_{kl}(i)w_{kl} / N$, we get an equivalent formulation. This formulation avoids the size differences in the K_i, but does penalize the alternative i whose criteria set K_i contains low ranked criteria versus an alternative that may be evaluated in terms of a similar number, but of higher ranked criteria. As with the previous model, there may be circumstances where this is a desirable property, and others where it is not.

Benevolent evaluation: performance relative to the ideal

In the case where we want to evaluate alternatives in the fairest possible (i.e., most *benevolent*) way, it can be argued that such an evaluation should not penalize an alternative for failing to be considered in terms of a large portion of the criteria, nor for failing to be evaluated relative to the most important criteria. This approach would then advocate evaluating an alternative in terms of only those criteria k on which it receives a ranking l. Only the importance of these "pertinent" criteria *relative to one another* would then come into play, and the standing of these criteria vis-à-vis the complementary set (the set on which i is not evaluated) would not enter the picture.

One means of accomplishing the aforementioned benevolent approach is to compare each alternative i to the best possible or ideal performance for that alternative. In the notation of the earlier model, the *ideal alternative* would receive a rating of

$$R_{\text{ideal}} = \sum_{k\in K} w_{kl} .$$

Clearly, any alternative i which ranks lower than first place ($l > 1$) on any criterion k will score worse than this ideal, hence $R_i \le R_{\text{ideal}}$. Thus, the measure

$$\hat{R}_i = R_i / R_{\text{ideal}} ,$$

is a reasonable and convenient way of expressing the performance level of i. $\hat{R}_i$ is similar in some respects to an industrial productivity measure where we compare actual to *standard* performance, although it could be argued that R_{standard} is probably something less than R_{ideal}. For our purposes, R_{ideal} represents the only tangible (and, in principle, achievable) measure that can be used as a backdrop against which to evaluate alternatives.

With this concept as a basis, and proceeding in a manner analogous to problem (7.7)-(7.13), consider the following N problems:

$$\hat{R}_{i_o}^* = \max \hat{R}_{i_o} = \frac{\sum_{k \in K_{i_o}} \sum_{l=1}^{L} d_{kl}(i_o) w_{kl}}{\sum_{k \in K_{i_o}} w_{kl}} \tag{7.17}$$

subject to

$$\frac{\sum_{k \in K_i} \sum_{l=1}^{L} d_{kl}(i) w_{kl}}{\sum_{k \in K_i} w_{k1}} \leq 1, \ i = 1,\ldots,N; \tag{7.18}$$

(7.8)-(7.13).

In this ratio formulation, the numerator in (7.18) represents the actual performance of alternative i, with the denominator being the theoretical or best possible performance. It is noted that in this formulation constraints (7.18) are redundant, and can, therefore, be removed from the problem. Unlike the linear problem, this formulation having a fractional objective function, is nonlinear, and in general can be difficult to solve. By way of a transformation, however, this problem can be converted to a linear format. Specifically, let

$$\tau_o = 1/\sum_{k \in K_{i_o}} w_{k1}$$

and define the variables $\tilde{w}_{kl} = \tau_o w_{kl}, \tilde{u}_k, = \tau_o u_k, \tilde{v} = \tau_o v$ and $\tilde{F} = \tau_o F$. Problem (7.7)-(7.13) (in the absence of constraints (7.18)) can then be written in the form:

$$\hat{R}_{i_o} = \max \hat{R}_{i_o} = \sum_{k \in K_{i_o}} \sum_{l=1}^{L} d_{kl}(i_o) \tilde{w}_{kl} \tag{7.19}$$

subject to

$$\sum_{k \in K_{i_o}} \tilde{w}_{kl} = 1; \tag{7.20}$$

$$\tilde{w}_{kl} - \tilde{w}_{kl+1} - g_l t_k \tilde{u}_k \geq 0,$$
$$k = 1,\ldots,K; \ l = 1,\ldots L-1;$$
$$\tilde{w}_{kL} - g_L t_k \tilde{u}_k \geq 0, \quad k = 1,\ldots,K; \tag{7.21a}$$

$$\tilde{w}_{kl} - \tilde{w}_{kL} - t_k(\tilde{w}_{11} - \tilde{w}_{1L}) \le 0,$$
$$k = 1,...,K; \qquad (7.21b)$$
$$\tilde{w}_{kl} - \tilde{w}_{k+1l} - \tilde{v}H_k \ge 0,$$
$$k = 1,...K-1;\ l = 1,...,L;$$
$$\tilde{w}_{Kl} - \tilde{v}H_K \ge 0,\ \ l = 1,...L; \qquad (7.22)$$
$$\tilde{u}_{\langle j\rangle} - \tilde{u}_{\langle j+1\rangle} - \tilde{F} \ge 0,\ \ j = 1,...,K-1; \qquad (7.23)$$
$$\tilde{\upsilon} \ge \tau_o z; \qquad (7.24)$$
$$\tilde{F} \ge \tau_o z; \qquad (7.25)$$
$$\tau_o, \tilde{w}_{kl}, \tilde{u}_k, \tilde{v}, \tilde{F} \ge 0, \quad \forall k,l. \qquad (7.26)$$

Note that this formulation is somewhat more elaborate that (7.7)-(7.13). The reader is referred to Cook et al. (1996).

Lemma 7.1: There exists an optimal solution to problem (7.17)-(7.18) in which $\sum_{k\in K_{i_o}} w_{k1} \le 1$.

Proof: For any feasible solution $W = (w_{kl})$ to (7.17)-(7.18), cW is also a feasible solution for any $c \ge 1$. Hence, we may impose a bounding constraint $\sum_{k\in K_{i_o}} w_{k1} \le \theta$ in (7.17)-(7.18) for some θ, and still have an equivalent problem. Furthermore, for z small enough we may, with no loss of generality, arbitrarily choose $\theta = 1$. Hence, the result. Q.E.D.

Lemma 7.2: There exists an optimal solution $\tilde{w}_{kl}^*, \tilde{u}_k^*, \tilde{v}^*, \tilde{F}^*, \tau_o^*$ to (7,19)-(7.26) in which $\tau_o^* = 1$.

Proof: Due to Lemma 7.1 and the definition of τ_o, we have

$$\tau_o^* = \frac{1}{\sum_{k\in K_{i_o}} w_{k1}^*} \ge 1.$$

To yield maximum flexibility in the problem, it is optimal to force τ_o to its lower limit (the problem is the least restricted in this case). Hence $\tau_o^* = 1$. Q.E.D.

Theorem 7.5: In the special case where all $K_i = K$ and $|K| = K$, problem (7.19)-(7.26) is equivalent to problem (7.7)-(7.13) if an $(N+1)$st alternative, the ideal alternative, is added to the latter.

Proof: From Lemma 7.2 $\tau_o^* = 1$, hence $w_{kl}^* = \tilde{w}_{kl}^*, u_k^* = \tilde{u}_k^*, v = \tilde{v}^*, F^* = \tilde{F}^*$. Furthermore, constraint (7.20) may be replaced by $\sum_{k=1}^{K} w_{k1} \le 1$, the upper limit on the rating for the ideal alternative. Since constraints (7.8) are redundant in the presence of this inequality, the result follows. Q.E.D.

By virtue of Theorem 7.5, problem (7.17)-(7.18) can be written in the form:

$$\hat{R}_{i_o} = \max\ R_{i_0} = \sum_{k \in K_{i_o}} \sum_{l=1}^{L} d_{kl}(i_o) w_{kl} \tag{7.27}$$

subject to

$$\sum_{k \in K_{i_o}} w_{kl} \le 1; \tag{7.28}$$

(7.9)-(7.13).

Common set of weights

As with problem (7.7)-(7.13), (7.27)-(7.28) will generally yield a different set of weights w_{kl} for each alternative i_o being evaluated. Along the lines of the previous section, a common set of weights can be derived by solving the problem:

$$z^* = \max\ z \tag{7.29}$$

subject to

$$\sum_{k \in K_i} w_{k1} \le 1, \quad i = 1, \ldots, N; \tag{7.30}$$

(7.8)-(7.13).

This problem is clearly bounded since every criterion k can be assumed to lie in at lease one subset K_i, hence $w_{kl} \le 1$ for all k. Thus, z will achieve an optimum. The final ratings to be assigned to any alternative i is given by

$$\tilde{R}_i = \frac{\sum_{k \in K_i} \sum_{l=1}^{L} d_{kl}(i) w_{kl}^*}{\sum_{k \in K_i} w_{k1}^*},$$

where the w_{kl}^* are the optimal variables from problem (7.29)-(7.30).

Model (7.17)-(7.18) (hence model (7.29)-(7.30)) has the advantage that it provides a fair evaluation to an alternative i, regardless of the status of those criteria K_i that pertain to that alternative. Specifically, an alternative is not penalized for or given an unfair advantage because of the nature of its particular criteria. This very property may in certain circumstances, however, be seen as a weakness of the approach. If in a project rating situation, for example, the contribution of projects to a specific management goal is a key element in deciding on the set of choices to be funded, then the model of this section may not be appropriate. On the other hand, if projects from different departments are to be fairly assessed so that all contenders have an opportunity to compete, then it may be desirable not to have criteria not pertinent to an alternative, affect how that alternative is rated in a relative sense.

7.5. CONCLUSIONS

This chapter has examined MCDM problems in the context of DEA. A multicriteria composite index model is presented which can accommodate both qualitative and quantitative data. Various example problem settings are given. The chapter also extends the methodology to handle situations where some alternatives may be evaluated only in terms of a proper subset of the full set of criteria. The proposed approach is based on examining performance of an alternative relative to an *ideal* status for that alternative.

REFERENCES

1. Ang, J., and T. Lai (1990), The discount rate for overseas investments under exchange and political risk, *In Pacific-Basin Capital Markets Research*, 467-483.
2. Blask, J. (1978), A survey of country evaluation systems in use, *Financing and Risk in Developing Countries*, Praeger, 65-78.
3. Charnes, A., W.W. Cooper and E. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2, 429-444.
4. Cook, W.D., J. Doyle, R. Green and M. Kress (1997), Multiple criteria modeling and ordinal data: Evaluation in terms of subsets of criteria, *European Journal of Operational Research* 98, 602-609.
5. Cook, W.D., and M. Kress (1994), A multiple-criteria composite index model for quantitative and qualitative data, *European Journal of Operational Research* 78, 367-379.
6. Cook, W.D., and K.J. Hebner (1993), A multicriteria approach to country risk evaluation: with an example employing Japanese data, *International Review of Economics and Finance* 2, 327-348.

7. Cook, W.D., and D.A. Johnston (1992), Evaluating suppliers of complex systems: a multiple criteria approach, *Journal of Operational Research Society* 43 (11), 1055-1061.
8. Cook, W.D., and M. Kress (1991), A multiple crisis decision model with ordinal preference data, *European Journal of Operational Research* 54, 191-198.
9. Fishbein, M (1967), A behavioral theory approach to the relations between beliefs about an object and the attitude towards the object, in M. Fishbein (ed.), *Readings in Attitude and Theory Measurement,* Wiley, New York, 389-399.
10. Goddard, S. (1990), Political risk in international capital budgeting, *Managerial Finance* 16, 7-12.
11. Hwang, C.-L., and K. Yoon (1981), Multiple attribute decision-making methods and applications, Lecture Notes in Economics and Mathematical Systems, no. 186, Springer-Verlag, Berlin.
12. Keeney, R.L., and H. Raiffa (1976), Decisions with multiple objectives, preferences and value tradeoffs, Wiley, New York.
13. Kobrin, S. (1986), Assessing political risk overseas, *The Multinational Enterprise in Transition*, 425-439.
14. Korhonen, P., and J. Wallenius (1990), Using qualitative data in multiple objective linear programming, *European Journal of Operational Research* 48, 81-87.
15. Oral, H., and R. Yolalan (1990), An empirical study of measuring operating efficiency and profitability of bank branches, *European Journal of Operational Research* 46, 282-294.
16. Parker, B.R., and V. Srinivasan (1976), A consumer preference approach to the planning of rural primary health care facilities, *Operations Research* 24, 991-1025.
17. Phillips-Patrick, F. (1989), The effect of asset and ownership structgure on political risk: evidence from Mitterrand's election in France, *Journal of Banking and Finance*, 651-671.
18. Phillips-Patrick, F. (1990), Ownership, asset structure, and political risk, *Financial Planning and Forecasting: International Dimensions of Securities and Currency Markets*, JAI Press, 239-256
19. Rosenberg, M.J. (1956), Cognitive structure and attitudinal affect, *Journal of Abnormal and Social Psychology,* 53, 367-372.
20. Saaty, T.L. (1980), The analytic hierarchy process, McGraw-Hill, New York.
21. Sherman, H.D., and F. Gold (1985), Bank branch operating efficiency: evaluation with data envelopment analysis, *Journal of Banking and Finance* 9, 297-315.

22. Srinivasan, V., and A.D. Shocker (1973), Estimating the weights for multiple attributes in a composite criterion using pairwise judgements, *Psychometrika* 38, 473-493.
23. Ting, W. (1988), Multinational risk assessment and management, *Quorum Books*.

This chapter is based upon (i) Cook, W.D., J. Doyle, R. Green and M. Kress, Multiple Criteria Modeling and Ordinal Data: Evaluation in Terms of Subsets of Criteria, European Journal of Operational Research, Vol. 89, 1996: pp. 602-609, with permission from Elsevier Science, (ii) Cook, W.D. and K. Hebner, A Multicriteria Approach to Country Risk Evaluation: With An Example Employing Japanese Data, International Review of Economics and Finance, Vol. 2, No. 4, 1993: pp. 327-348, with permission from International Review of Economics and Finance, and (iii) Cook, W.D. and D. Johnston, Evaluating Suppliers of Complex Systems: A Multiple Criteria Approach, Journal of the Operational Research Society, Vol. 43, No. 11, 1992: pp. 1055-1061, with permission from Palgrave.

Chapter 8

MODELING RANK ORDER DATA

8.1. INTRODUCTION

In a wide range of problem settings to which DEA can be applied, particularly in not-for-profit cases, *qualitative* factors are often present. In some situations such factors may be legitimately "quantifiable," but very often such quantification is superficially forced, as a modeling convenience. Typically, a qualitative factor such as management competence, for example, is captured either on a Likert scale, or is represented by some quantitative *surrogate* such as plant downtime or percentage sick days by employees.

It can be the case as well, that purely *quantitative* variables may be such that accurate data is not available, hence figures provided are often rough estimates of the actual data values. In a number of studies of bank and bank branch efficiency, for example, discretionary inputs such as "percentage of high value customers" in the customer base, can be an important influence variable vis-à-vis performance. It reflects investment potential on the part of the customer. See Cook, Hababou and Tuenter (2000) and Cook and Hababou (2001). This variable is, however, generated from disposable income of the customer, for which accurate data is seldom available. For existing branches, a surrogate for such a variable is the level of investment of the customer. For new (planned) branches, the level of investment that would be created can be predicted from income demographics for the customer base for that branch. Such income data is, however, often unreliable.

In situations such as those described, the "data" for certain influence factors (inputs and outputs) might better be represented as rank positions in an ordinal, rather than numerical sense. Refer again to the management competence example. In certain circumstances, the information available may permit one *only* to put each DMU into one of L categories or groups (e.g. 'high', 'medium' and 'low' competence). In other cases, one may be able to provide a complete rank ordering of the DMUs on such a factor.

This chapter examines the modeling of qualitative data in the DEA structure. The following Section 8.2 discusses two practical problem settings in which qualitative data occurs naturally. In the first, we examine a problem of R&D project ranking and selection, where various non-quantifiable factors need to be considered. In the context of DEA, the projects represent the decision making units. This example is adopted from Cook et al. (1996). In the second example, due to Kim et al. (1999), and Zhu (2003), a mix of ordinal and numerical factors are evaluated. Section 3 examines the radial projection DEA model in the context of ordinal data. Section 4 discusses the application of this ordinal DEA model to the two presented problems. In Section 5, various settings involving ordinal data are discussed. Conclusions and further directions are presented in Section 6.

8.2. PROBLEM SETTINGS INVOLVING ORDINAL DATA

8.2.1 Ordinal Data in R&D Project Selection

Consider the problem of selecting R&D projects in a major public utility corporation with a large research and development branch. Research activities are housed within several different divisions, for example, thermal, nuclear, electrical, and so on. In a budget constrained environment in which such an organization finds itself, it becomes necessary to make choices among a set of potential research initiatives or projects that are in competition for the limited resources. To evaluate the impact of funding (or not funding) any given research initiative, two major considerations generally must be made. First, the initiative must be viewed in terms of more than one factor or criterion. Second, some or all of the criteria that enter the evaluation may be qualitative in nature. Even when clearly quantitative factors are involved, such as long term savings to the organization, it may be extremely difficult to obtain even a crude estimate of the value of that factor. The most that one can do in many such situations is to classify the project (according to this factor) on some scale (high/medium/low or say a 5-point scale).

Let us assume that for each qualitative criterion, each initiative is rated on a 5-point scale, where the particular point on the scale is chosen through a consensus on the part of executives within the organization. Table 8-1 presents an illustration of how the data might appear for 10 projects, 3 qualitative output criteria (benefits), identified as 1, 2, and 3, and 3 qualitative input criteria (cost of resources), identified as 4,5, and 6. In the actual setting examined, a number of potential benefit and cost criteria were considered as displayed in Tables 8-2 and 8-3.

We use the convention that for both outputs and inputs, a rating of 1 is "best", and 5 "worst". For outputs, this means that a DMU ranked at position 1 generates *more* output than is true of a DMU in position 2, and so on. For inputs, a DMU in position 1 consumes *less* input than one in position 2.

Table 8-1. Ratings by Criteria

Project No.	Outputs			Inputs		
	1	2	3	4	5	6
1	2	4	1	5	2	1
2	1	1	4	3	5	2
3	1	1	1	1	2	1
4	3	3	3	4	3	2
5	4	3	5	5	1	4
6	2	5	1	1	2	2
7	1	4	1	5	4	3
8	1	5	3	3	3	3
9	5	2	4	4	2	5
10	5	4	4	5	5	5

Regardless of the manner in which such a scale rating is arrived at, the conventional DEA model is capable only of treating the information as if it has cardinal meaning (e.g. something which receives a score of 4 is evaluated as being twice as important as something that scores 2). There are a number of problems with this approach. First and foremost, the projects' original data in the case of some criteria may take the form of an ordinal ranking of the projects. Specifically, the most that can be said about two projects i and j is that i is preferred to j. In other cases it may only be possible to classify projects as say 'high', 'medium' or 'low' in importance on certain criteria. When projects are rated on, say, a 5-point scale, it is generally understood that this scale merely provides a relative positioning of the projects. In a number of agencies investigated (for example, hydro electric and telecommunications companies), 5-point scales are common for evaluating alternatives in terms of qualitative data, and are often accompanied by interpretations such as:

1 = Extremely important
2 = Very important
3 = Important
4 = Low in importance
5 = Not important,

which are easily understood by management. While it is true that market researchers often treat such scales in a numerical (i.e. cardinal) sense, no one seriously believes that an 'extremely important' classification for a project should be interpreted literally as meaning that this project rates three times better than one which is only classified as 'important.' The key message here is that many, if not all criteria, used to evaluate R&D projects are qualitative in nature, and should be treated as such. The model presented in the following sections extends the DEA idea to an ordinal setting, hence accommodating this very practical consideration.

Table 8-2. Potential Benefits

Criteria	Sub-criteria or Interpretation
1. Enhancement of energy efficiency	-development of high yield technologies -initiatives which will reduce energy demand -development of technologies for utilizing residues
2. Enhancement of diversification/alternative energy sources	-initiatives which provide or strive for new energy sources -provide for flexibility in or adaptability of existing and new facilities
3. $Saved internal to organization	-cost reduction devices -new technology to replace obsolete equipment
4. Impact on environment	-reduction of emissions into water and atmosphere -reduction of risk of nuclear accidents
5. Enhancement to internal technical capability and research profile	- provides training and develops expertise -provides technical resources (software, equipment, etc.) -builds linkages to external research community.
6. Enhancement to research profile as viewed by the external community	-impact on research status among other utility companies -impact on profile abroad
7. Economic impact on external community	-job creation outside organization -$ savings to public and industry created by energy efficiency devices
8. Impact on nuclear performance	-influence on nuclear station maintenance, etc.

Table 8-3. Potential Costs

Criteria	Sub-criteria or Interpretation
1. Technical expertise available internally	
2. Technical expertise available externally	consultants
	other research centres
3. Technology available	equipment
	software

8.2.2 Efficiency Performance of Korean Telephone Offices

Kim et al. (1999) examine 33 telephone offices in Korea and use the following factors to develop performance measures.

Inputs
(1) manpower
(2) operating costs
(3) number of telephone lines

Outputs
(1) local revenues
(2) long distance revenues
(3) international revenues
(4) operation/maintenance level
(5) customer satisfaction.

All inputs and outputs (1),(2),(3) are quantitative, and can be used in the DEA framework in the usual way. Output #4 is, however, ordinal and provides a complete ranking of the 33 DMUs. Output #5 is a categorization of the DMUs on a 5-point Likert scale. Table 8-4 displays the data.

In the section to follow the conventional DEA structure is adapted to accommodate variables measured on an ordinal scale.

8.3. MODELING ORDINAL DATA

The above problems typify situations in which pure ordinal data, or a mix of ordinal and numerical data, are involved in the performance measurement exercise. There appear to be two general approaches in the literature to the handling of ordinal/qualitative data within the DEA framework. The first effort was presented in Cook et al. (1993), (1996). The general approach given below leads ultimately to their model. The second and related effort is that due to Cooper et al. (1999), under the title *imprecise data.* Again, using the general structure given below, one arrives at their model. Rather than adopting, outright, one or the other of these approaches, let us cast the ordinal data problem in a general DEA format. Specifically, consider the situation in which a set of N decision making units (DMUs), k=1,...N are to be evaluated in terms of R_1 numerical outputs, R_2 ordinal outputs, I_1 numerical inputs, and I_2 ordinal inputs. Let $Y_k^1 = (y_{rk}^1)$, $Y_k^2 = (y_{rk}^2)$ denote the R_1- dimensional and R_2-dimensional vectors of outputs, respectively.

Table 8-4. Data for Telephone Offices

DMU No	X1	X2	X3	Y1	Y2	Y3	Y4	Y5
1	239	7.03	158	47.1	16.67	34	28	2
2	261	3.94	163	37.5	14.11	20	26	3
3	170	2.1	90	20.7	6.8	12.6	19	3
4	290	4.54	201	41.8	11.07	6.27	23	4
5	200	3.99	140	33.4	9.81	6.49	30	2
6	283	4.65	214	42.4	11.34	5.16	21	4
7	286	6.54	197	47	14.62	13	9	2
8	375	6.22	314	55.5	16.39	7.31	14	1
9	301	4.82	257	49.2	16.15	6.33	8	3
10	333	6.87	235	47.1	13.86	6.51	6	2
11	346	6.46	244	49.4	15.88	8.87	18	2
12	175	2.06	112	20.4	4.95	1.67	32	5
13	217	4.11	131	29.4	11.39	4.38	33	2
14	441	7.71	214	61.2	25.59	33	16	3
15	204	3.64	163	32.3	9.57	3.65	15	4
16	216	2.24	154	32.8	11.46	9.02	25	2
17	347	5.65	301	59	17.82	8.19	29	1
18	288	4.66	212	42.3	14.52	7.33	24	4
19	185	3.37	178	33	9.46	2.91	7	2
20	242	5.12	270	65.1	24.57	20.7	17	1
21	234	2.52	126	31.6	8.55	7.27	27	2
22	204	4.24	174	32.5	11.15	2.95	22	3
23	356	7.95	299	66	22.25	14.9	13	2
24	292	4.52	236	50	14.77	6.35	12	3
25	141	5.21	63	21.5	9.76	16.3	11	2
26	220	6.09	179	47.9	17.25	22.1	31	2
27	298	3.44	225	42.4	11.14	4.25	4	2
28	261	4.3	213	41.7	11.13	4.68	20	5
29	216	3.86	156	31.6	11.89	10.5	3	3
30	171	2.45	150	24.1	9.08	2.6	10	5
31	123	1.72	61	12	4.78	2.95	5	1
32	89	0.88	42	6.4	3.18	1.48	2	5
33	109	1.35	57	10.6	3.43	2	1	4

Similarly, let $X_k^1 = (x_{ik}^1)$ and $X_k^2 = (x_{ik}^2)$ be the I_1-dimensional and I_2-dimensional vectors of inputs, respectively.

In the situation where all factors are quantitative, the conventional radial projection model for measuring DMU efficiency is expressed by the ratio of weighted outputs to weighted inputs. Adopting the general variable returns to scale (VRS) model of Banker, Charnes and Cooper (1984), and stating it in ratio form, the efficiency of DMU "o" follows from the solution of:

$$
\begin{aligned}
& e_o = \max \left(\mu_o + \sum_{r\in R_1} \mu_r^1 y_{ro}^1 + \sum_{r\in R_2} \mu_r^2 y_{ro}^2\right) / \\
& \left(\sum_{i\in I_1} \upsilon_i^1 x_{io}^1 + \sum_{i\in I_2} \upsilon_i^2 x_{io}^2\right) \\
& \text{subject to:} \\
& \left(\mu_o + \sum_{r\in R_1} \mu_r^1 y_{rk}^1 + \sum_{r\in R_2} \mu_r^2 y_{rk}^2\right) / \\
& \left(\sum_{i\in I_1} \upsilon_i^1 x_{ik}^1 + \sum_{i\in I_2} \upsilon_i^2 x_{ik}^2\right) \leq 1, \text{ all k} \\
& \mu_r^1,\ \mu_r^2,\ \upsilon_i^1, \upsilon_i^2 \geq \varepsilon, \text{ all r, i}
\end{aligned}
\tag{8.1}
$$

Problem (8.1) is convertible to the linear programming format:

$$
\begin{aligned}
& e_o = \max\ \mu_o + \sum_{r\in R_1} \mu_r^1 y_{ro}^1 + \sum_{r\in R_2} \mu_r^2 y_{ro}^2 \\
& \text{subject to} \\
& \sum_{i\in I_1} \upsilon_i^1 x_{io}^1 + \sum_{i\in I_2} \upsilon_i^2 x_{io}^2 = 1 \\
& \mu_o + \sum_{r\in R_1} \mu_r^1 y_{rk}^1 + \sum_{r\in R_2} \mu_r^2 y_{rk}^2 - \\
& \sum_{i\in I_1} \upsilon_i^1 x_{ik}^1 - \sum_{i\in I_2} \upsilon_i^2 x_{ik}^2 \leq 0, \text{ all k} \\
& \mu_r^1,\ \mu_r^2,\ \upsilon_i^1,\ \upsilon_i^2 \geq \varepsilon, \text{ all r, i,}
\end{aligned}
\tag{8.2}
$$

whose dual is given by

$$
\begin{aligned}
& \min\ \theta - \varepsilon \sum_{r\in R_1 U R_2} s_r^+ - \varepsilon \sum_{i\in I_1 U I_2} s_i^- \\
& \text{subject to} \\
& \sum_{k=1}^{N} \lambda_k y_{rk}^1 - s_r^+ = y_{ro}^1,\ r \in R_1 \\
& \sum_{n=1}^{N} \lambda_k y_{rk}^2 - s_r^+ = y_{ro}^2,\ r \in R_2 \\
& \theta x_{io}^1 - \sum_{k=1}^{N} \lambda_k x_{ik}^1 - s_i^- = 0, i \in I_1 \\
& \theta x_{io}^2 - \sum_{k=1}^{N} \lambda_k x_{ik}^2 - s_i^- = 0,\ i \in I_2 \\
& \sum_{k=1}^{N} \lambda_k = 1 \\
& \lambda_k, s_r^+, s_i^- \geq 0, \text{ all k,r, i }, \theta \text{ unrestricted}
\end{aligned}
\tag{8.2'}
$$

For the problem settings described in the previous section, precise values for outputs in R_2 and inputs in I_2 are not available. Cooper et al. (1999), (2001), and Zhu (2003) refer to this as an example of *imprecise* DEA or IDEA. To place the problem in a general framework, assume that for each ordinal factor ($r \in R_2$, $i \in I_2$), a DMU k can be assigned to one of L rank positions, where $L \leq N$. As discussed earlier, L=5, is an example of an appropriate number of rank positions in many practical situations. We point out that in certain application settings, different ordinal factors may have different L-values associated with them. For example, in the problem described in subsection 2.2, 'customer satisfaction,' y_5 is measured on a 5-point scale, while 'operation/maintenance level,' y_4 provides for a full ranking of all 33 DMUs (L=33). For exposition purposes, we assume a common L-value throughout. We demonstrate later that this provides no loss of generality.

In the development below it is assumed that a "full ranking" of all DMUs is available for each ordinal factor. That is, each DMU is assumed to occupy a rank position on each ordinal factor, as opposed to there being only a *partial ranking* of the DMUs on some factor. In Section 5 we discuss a situation where such partial ranking does occur.

One can view the allocation of a DMU to a rank position ℓ on an output r, for example, as having assigned that DMU an output *value* or *worth* $y_r^2(\ell)$. The implementation of the DEA model (8.1) (and (8.2)) thus involves determining two things:

(1) multiplier values μ_r^2, υ_i^2 for outputs $r \in R_2$ and inputs $i \in I_2$;
(2) rank position values $y_r^2(\ell)$, $r \in R_2$, and $x_i^2(\ell)$, $i \in I_2$, all ℓ.

Cooper et al. (1999) use a similar format to the one presented here, and approach this problem in a two-stage manner. Their approach for handing *imprecise data* first derives appropriate values (in our notation) for the $y_r^2(\ell)$ and $x_i^2(\ell)$ (i.e., they resolve item (2) above). These values having now been quantified, the conventional DEA model (8.2) can be solved. In this section we show that the problem can be reduced to the standard VRS model by considering items (1) and (2) simultaneously. Further mention of IDEA appears later.

To facilitate development herein, define the L-dimensional unit vectors $\gamma_{rk} = (\gamma_{rk}(\ell))$, and $\delta_{ik} = (\delta_{ik}(\ell))$ where

$$\gamma_{rk}(\ell) = \begin{cases} 1 & \text{if DMU k is ranked in } \ell \text{ th position on output r} \\ 0, & \text{otherwise} \end{cases}$$

$$\delta_{ik}(\ell) = \begin{cases} 1 & \text{if DMU k is ranked in } \ell \text{ th position on input i} \\ 0, & \text{otherwise} \end{cases}$$

For example, if a 5-point scale is used, and if DMU #1 is ranked in ℓ = 3rd place on ordinal output r=5, then $\gamma_{51}(3) = 1$, $\gamma_{51}(\ell) = 0$, for all other rank positions ℓ. Thus, y^2_{51} is assigned the value $y^2_5(3)$, the *worth* to be credited to the 3rd rank position on output factor 5. It is noted that y^2_{rk} can be represented in the form

$$y^2_{rk} = y^2_r(\ell_{rk}) = \sum_{\ell=1}^{L} y^2_r(\ell)\, \gamma_{rk}(\ell),$$

where ℓ_{rk} is the rank position occupied by DMU k on output r. Hence, model (8.2) can be rewritten in the more representative format.

$$e_o = \max\ \mu_o + \sum_{r\in R_1} \mu^1_r y^1_{ro} + \sum_{r\in R_2} \sum_{\ell=1}^{L} \mu^2_r y^2_r(\ell)\, \gamma_{ro}(\ell)$$

subject to:

$$\sum_{i\in I_1} \upsilon^1_i x^1_{io} + \sum_{i\in I_2} \sum_{\ell=1}^{L} \upsilon^2_i x^2_i(\ell)\, \delta_{io}(\ell) = 1$$

$$\mu_o + \sum_{r\in R_1}\mu^1_r y^1_{rk} + \sum_{r\in R_2}\sum_{\ell=1}^{L}\mu^2_r y^2_r(\ell)\, \gamma_{rk}(\ell) - \sum_{i\in I_1}\upsilon^1_i x^1_{ik} -$$

$$\sum_{i\in I_2}\sum_{\ell=1}^{L}\upsilon^2_i x^2_i(\ell)\, \delta_{ik}(\ell) \le 0, \text{ all k} \qquad (8.3)$$

$$\{Y^2_r = (y^2_r(\ell)), X^2_i = (x^2_i(\ell))\} \in \Psi$$

$$\mu^1_r,\ \upsilon^1_i \ge \varepsilon$$

In (8.3) we use the notation Ψ to denote the set of *permissible worth vectors*. We discuss this set below.

It must be noted that the same infinitesimal ε is applied here for the various input and output multipliers, which may, in fact, be measured on scales that are very different from another. If two inputs are, for example, x^1_{i1k} representing 'labor hours', and x^1_{i2k} representing 'available computer technology', the scales would clearly be incompatible. Hence, the likely sizes of the corresponding multipliers υ^1_{i1}, υ^1_{i2} may be similarly different. Thrall (1996) has suggested a mechanism for correcting for such scale incompatibility, by applying a *penalty vector* G to augment ε, thereby creating differential lower bounds on the various υ_i, μ_r. Proper choice of G can effectively bring all factors to some form of common scale or unit. For simplicity of presentation we will assume the cardinal scales for all $r \in R_1$, $i \in I_1$ are similar in dimension, and that G is the unit vector. The more general case would proceed in an analogous fashion.

Permissible Worth Vectors

The values or worths $\{y_r^2(\ell)\}$, $\{x_i^2(\ell)\}$, attached to the ordinal rank positions for outputs r and inputs i, respectively, must satisfy the minimal requirement that it is *more* important to be ranked in ℓ^{th} position than in the $(\ell+1)^{st}$ position on any such ordinal factor. Specifically, $y_r^2(\ell) > y_r^2(\ell+1)$ and $x_i^2(\ell) < x_i^2(\ell+1)$. That is, for outputs, one places a higher weight on being ranked in ℓ^{th} place than in $(\ell+1)^{st}$ place. For inputs, the opposite is true. A set of linear conditions that produce this realization is defined by the set Ψ, where

$$\Psi = \{(Y_r^2, X_i^2) | \; y_r^2(\ell) - y_r^2(\ell+1) \geq \varepsilon, \; \ell=1, \ldots L-1, \; y_r^2(L) \geq \varepsilon, \; x_i^2(\ell+1) - x_i^2(\ell) \geq \varepsilon, \; \ell=1, \ldots L-1, \; x_i^2(1) \geq \varepsilon\}.$$

Arguably, ε could be made dependent upon ℓ (i.e. replace ε by ε_l). It can be shown, however, that all results discussed below would still follow. For convenience, we, therefore, assume a common value for ε.

We now demonstrate that the nonlinear problem (8.3) can be written as a linear programming problem.

Theorem 8.1
Problem (8.3), in the presence of the permissible worth space Ψ, can be expressed as a linear programming problem.

Proof: In (8.3), make the change of variables

$$w_{r\ell}^1 = \mu_r^2 y_r^2(\ell), \; w_{i\ell}^2 = \upsilon_i^2 x_i^2(\ell)$$

It is noted that in Ψ, the expressions

$$y_r^2(\ell) - y_r^2(\ell+1) \geq \varepsilon, \; y_r^2(L) \geq \varepsilon$$

can be replaced by

$$\mu_r^2 y_r^2(\ell) - \mu_r^2 y_r^2(\ell+1) \geq \mu_r^2 \varepsilon, \; \mu_r^2 y_r^2(L) \geq \mu_r^2 \varepsilon,$$

which becomes

$$w_{r\ell}^1 - w_{r\ell+1}^1 \geq \mu_r^2 \varepsilon, \; w_{rL}^2 \geq \mu_r^2 \varepsilon.$$

A similar conversion holds for the $x_i^2(\ell)$.

Problem (8.3) now becomes

$$e_o = \max \mu_o + \sum_{r\in R1} \mu_r^1 y_{ro}^1 + \sum_{r\in R2} \sum_{\ell=1}^{L} w_{r\ell}^1 \gamma_{ro}(\ell)$$

subject to

$$\sum_{i\in I1} \upsilon_i^1 x_{io}^1 + \sum_{i\in I2} \sum_{\ell=1}^{L} w_{i\ell}^2 \delta_{io}(\ell) = 1$$

$$\mu_o + \sum_{r\in R1} \mu_r^1 y_{rk}^1 + \sum_{r\in R2} \sum_{\ell=1}^{L} w_{rl}^1 \gamma_{rk}(\ell) - \sum_{i\in I1} \upsilon_i^1 x_{ik}^1 - \sum_{i\in I2} \sum_{\ell=1}^{L} w_{i\ell}^2 \delta_{ik}(\ell) \leq 0, \text{ all k} \quad (8.4)$$

$$w^1_{r\ell} - w^1_{r\ell+1} \geq \mu^2_r\, \varepsilon\,,\ \ell=1,\ldots L-1, \text{ all } r \in R_2$$

$$w^1_{rL} \geq \mu^2_r\ \varepsilon\,, \text{ all } r \in R_2$$

$$w^2_{i\ell+1} - w^2_{i\ell} \geq \upsilon^2_i\ \varepsilon\,,\ \ell=1,\ \ldots \text{ L-1, all } i \in I_2$$

$$w^2_{i1} \geq \upsilon^2_i\ \varepsilon\,, \text{ all } i \in I_2$$

$$\mu^1_r,\ \upsilon^1_i \geq \varepsilon\,, \text{ all } r \in R_1, i \in I_1$$

$$\mu^2_r,\ \upsilon^2_i \geq \varepsilon\,, \text{ all } r \in R_2, i \in I_2$$

Problem (8.4) is clearly in linear programming problem format.

We state without proof the following theorem.

Theorem 8.2
At the optimal solution to (8.4), $\mu^2_r = \upsilon^2_i = \varepsilon$ for all $r \in R_2$, $i \in I_2$.

Problem (8.4) can then be expressed in the form:

$$e_o = \max\ \mu_o + \sum_{r\in R_1} \mu^1_r\, y^1_{ro} + \sum_{r\in R2} \sum_{\ell=1}^{L} w^1_{r\ell}\, \gamma_{ro}(\ell)$$

subject to

$$\sum_{i\in I1} \upsilon^1_i\, x^1_{io} + \sum_{i\in I2} \sum_{\ell=1}^{L} w^2_{i\ell}\, \delta_{io}(\ell) = 1$$

$$\mu_o + \sum_{r\in R1} \mu^1_r\, y^1_{rk} + \sum_{r\in R2} \sum_{\ell=1}^{L} w^1_{r\ell}\, \gamma_{rk}(\ell) -$$

$$\sum_{i\in I1} \upsilon^1_i\, x^1_{ik} - \sum_{i\in I2} \sum_{\ell=1}^{L} w^2_{i\ell}\, \delta_{ik}(\ell) \leq 0, \text{ all } k \tag{8.5}$$

$$-w^1_{r\ell} + w^1_{r\ell+1} \leq -\varepsilon^2,\ \ell=1,\ldots \text{L-1, all } r \in R_2$$

$$-w^1_{rL} \leq -\varepsilon^2, \text{ all } r \in R_2$$

$$-w^2_{i\ell+1} + w^2_{i\ell} \leq -\varepsilon^2,\ \ell=1, \ldots, \text{L-1, all } i \in I_2$$

$$-w^2_{i1} \leq -\varepsilon^2, \text{ all } i \in I_2$$

$$\mu^1_r,\ \upsilon^1_i \geq \varepsilon\,, r \in R_1, i \in I_1$$

It can be shown that (8.5) is equivalent to the standard VRS model. First we form the dual of (8.5).

$$\min\ \theta - \varepsilon \sum_{r\in R1} s_r^+ - \varepsilon \sum_{i\in I1} s_i^- - \varepsilon^2 \sum_{r\in R2}\sum_{\ell=1}^{L} \alpha_{r\ell}^1 - \varepsilon^2 \sum_{i\in I2}\sum_{\ell=1}^{L} \alpha_{i\ell}^2$$

subject to:

$$\sum_{k=1}^{N} \lambda_k y_{rk}^1 - s_r^+ = y_{ro}^1,\ r \in R_1$$

$$\theta x_{io}^1 - \sum_{k=1}^{N} \lambda_k x_{ik}^1 - s_i^- = 0,\ i \in I_1$$

$$\left.\begin{aligned} &\sum_{k=1}^{N} \lambda_k \gamma_{rk}(1) - \alpha_{r1}^1 = \gamma_{ro}(1) \\ &\sum_{k=1}^{N} \lambda_k \gamma_{rk}(2) + \alpha_{r1}^1 - \alpha_{r2}^1 = \gamma_{ro}(2) \\ &\vdots \\ &\sum_{k=1}^{N} \lambda_k \gamma_{rk}(L) + \alpha_{rL-1}^1 - \alpha_{rL}^1 = \gamma_{ro}(L) \end{aligned}\right\} r \in R_2 \tag{8.5'}$$

$$\left.\begin{aligned} &\delta_{io}(L)\,\theta - \sum_{k=1}^{N} \lambda_k \delta_{ik}(L) - \alpha_{iL}^2 = 0 \\ &\delta_{io}(L-1)\theta - \sum_{k=1}^{N} \lambda_k \delta_{ik}(L-1) + \alpha_{iL}^2 - \alpha_{iL-1}^2 = 0 \\ &\vdots \\ &\delta_{io}(1)\,\theta - \sum_{k=1}^{N} \lambda_k \delta_{ik}(1) + \alpha_{i2}^2 - \alpha_{i1}^2 = 0 \end{aligned}\right\} i \in I_2$$

$$\sum_{k=1}^{N} \lambda_k = 1$$

$$\lambda_k, s_r^+, s_i^-, \alpha_{r\ell}^1, \alpha_{i\ell}^2 \geq 0$$

θ unrestricted.

Here, we use $\{\lambda_k\}$ as the standard dual variables associated with the N ratio constraints, and the variables $\{\alpha_{i\ell}^2,\ \alpha_{r\ell}^1\}$ are the dual variables associated with the rank order constraints defined by Ψ. The slack variables s_r^+, s_i^- correspond to the lower bound restrictions on μ_r^1, υ_i^1.

Now, perform simple row operations on (8.5') by replacing the ℓ^{th} constraint by the sum of the first ℓ constraints. That is, the second constraint (for those $r \in R_2$ and $i \in I_2$) is replaced by the sum of the first two constraints, constraint 3 by the sum of the first three, and so on. Letting

$$\bar{\gamma}_{rk}(\ell) = \sum_{n=1}^{\ell} \gamma_{rk}(n) = \gamma_{rk}(1) + \gamma_{rk}(2) + \ldots + \gamma_{rk}(\ell),$$

and

$$\bar{\delta}_{ik}(\ell) = \sum_{n=\ell}^{L} \delta_{ik}(n) = \delta_{ik}(L) + \delta_{ik}(L-1) + \ldots + \delta_{ik}(\ell),$$

problem (8.5') can be rewritten as:

$$\begin{aligned}
&\min \theta - \varepsilon \sum_{r\in R1} s_r^+ - \varepsilon \sum_{i\in I1} s_i^- - \\
&\varepsilon^2 \sum_{r\in R2} \sum_{\ell=1}^{L} \alpha_{r\ell}^1 - \varepsilon^2 \sum_{i\in I2} \sum_{\ell=1}^{L} \alpha_{i\ell}^2 \\
&\text{subject to} \\
&\sum_{k=1}^{N} \lambda_k y_{rk}^1 - s_r^+ = y_{ro}^1, \ r \in R_1 \\
&\theta x_{io}^1 - \sum_{k=1}^{N} \lambda_k x_{ik}^1 - s_i^- = 0, \ i \in I_1 \qquad (8.6') \\
&\sum_{k=1}^{N} \lambda_k \bar{\gamma}_{rk}(\ell) - \alpha_{r\ell}^1 = \bar{\gamma}_{ro}(\ell), \\
&r \in R_2, \ \ell = 1, \ldots, L \\
&\theta \bar{\delta}_{io}(\ell) - \sum_{k=1}^{N} \lambda_k \bar{\delta}_{ik}(\ell) - \alpha_{il}^2 = 0, \ i \in I_2, \ \ell = 1, \ldots L \\
&\sum_{k=1}^{N} \lambda_k = 1 \\
&\lambda_k, s_r^+, s_i^-, \alpha_{r\ell}^1, \alpha_{i\ell}^2 \geq 0, \text{ all } i, r, \ell, k, \\
&\theta \text{ unrestricted in sign.}
\end{aligned}$$

The dual of (8.6') has the VRS format:

$$\begin{aligned}
&e_o = \max \mu_o + \sum_{r\in R1} \mu_r^1 y_{ro}^1 + \sum_{r\in R2} \sum_{\ell=1}^{L} w_{r\ell}^1 \bar{\gamma}_{ro}(\ell) \\
&\text{subject to} \\
&\sum_{i\in I1} \upsilon_i^1 x_{io}^1 + \sum_{i\in I2} \sum_{\ell=1}^{L} w_{i\ell}^2 \bar{\delta}_{io}(\ell) = 1 \qquad (8.6) \\
&\mu_o + \sum_{r\in R1} \mu_r^1 y_{rk} + \sum_{r\in R2} \sum_{\ell=1}^{L} w_{r\ell}^1 \bar{\gamma}_{rk}(\ell) - \sum_{i\in I1} \upsilon_i^1 x_{ik}^1 - \\
&\sum_{i\in I_2} \sum_{\ell=1}^{L} w_{i\ell}^2 \bar{\delta}_{ik}(\ell) \leq 0, \text{ all } k \\
&\mu_r^1, \upsilon_i^1 \geq \varepsilon, \ w_{r\ell}^1, w_{i\ell}^2 \geq \varepsilon^2,
\end{aligned}$$

which is a form of the VRS model. The slight difference between (8.6) and the conventional VRS model of Banker et al. (1984), is the presence of a different ε (i.e., ε^2) relating to the multipliers $w^1_{r\ell}$, $w^2_{i\ell}$, than is true for the multipliers μ^1_r, υ^1_i. It is observed that in (8.6') the common L-value can easily be replaced by criteria specific values (e.g. L_r for output criterion r). The model structure remains the same, as does that of model (8.6). Of course, since the intention is to have an infinitesimal lower bound on multipliers (i.e., ε >0), one can, from the start, restrict

$$\mu^1_r,\ \upsilon^1_i \geq \varepsilon^2$$

and

$$\mu^2_r,\ \upsilon^2_i \geq \varepsilon\,.$$

This leads to a form of (8.6) where all multipliers have the same infinitesimal lower bounds, making (8.6) precisely a VRS model in the spirit of Banker et a. (1984).

It is interesting to note that the IDEA approach of Cooper et al (1999) essentially involves tackling problem (8.2) by first attributing values to the imprecise data (rank positions), and second, optimizing (in the DEA structure) to arrive at optimal multipliers. The Cook et al (1993), (1996) approach to (8.2) is somewhat the reverse of this. It amounts ultimately to attributing values to the multipliers, and then letting the DEA optimization derive the values for the rank positions. Thus, these seemingly quite different approaches would appear to arrive at approximately the same final point.

Criteria Importance

The presence of ordinal data factors results in the need to *impute* values $y^2_r(\ell)$, $x^2_i(\ell)$ to outputs and inputs, respectively, for DMUs that are ranked at positions ℓ on an L-point Likert or ordinal scale. Specifically, all DMUs ranked at that position will be credited with the same "amount" $y^2_r(\ell)$ of output r ($r \in R_2$) and $x^2_i(\ell)$ of input i ($i \in I_2$).

A consequence of the change of variables undertaken above, to bring about linearization of the otherwise nonlinear terms, e.g., $w^1_{r\ell} = \mu^2_r y^2_r(\ell)$, is that at the optimum, all $\mu^2_r = \varepsilon^2$, $\upsilon^2_i = \varepsilon^2$. Thus, all of the ordinal criteria are relegated to the status of being of *equal importance*. Arguably, in many situations, one may wish to view the relative importance of these ordinal criteria (as captured by the μ^2_r, υ^2_i) in the same spirit as we have viewed the data values $\{y^2_{rk}\}$. That is, there may be sufficient information to be able to *rank* these criteria. Specifically, suppose that the R_2 output criteria can be grouped into L_1 categories and the I_2 input criteria into L_2 categories.

Now, replace the variables μ^2_r by $\mu^2(m)$, and υ^2_i by $\upsilon^2(n)$, and restrict:

$$\mu^2(m) - \mu^2(m+1) \geq \varepsilon, \ m=1,\ldots L_1-1$$
$$\mu^2(L_1) \geq \varepsilon$$

and

$$\upsilon^2(n) - \upsilon^2(n+1) \geq \varepsilon, \ n=1,\ldots,L_2-1$$
$$\upsilon^2(L_2) \geq \varepsilon.$$

Letting m_r denote the rank position occupied by output $r \in R_2$, and n_i the rank position occupied by input $i \in I_2$, we define the change of variables

$$w^1_{r\ell} = \mu^2(m_r)\, y^2_r(\ell)$$
$$w^2_{i\ell} = \upsilon^2(n_i)\, x^2_i(\ell)$$

The corresponding version of model (8.4) would see the lower bound restrictions μ^2_r, $\upsilon^2_i \geq \varepsilon$ replaced by the above constraints on $\mu^2(m)$ and $\upsilon^2(n)$. Again, arguing that at the optimum in (8.4), these variables will be forced to their lowest levels, the resulting values of the $\mu^2(m)$, $\upsilon^2(n)$ will be

$$\mu^2(m) = (L_1+1-m)\varepsilon, \ \upsilon^2(n) = (L_2+1-n)\varepsilon.$$

This implies that the lower bound restrictions on $w^1_{r\ell}$, $w^2_{i\ell}$ become

$$w^1_{r\ell} \geq (L_1+1-m_r)\ \varepsilon^2, \ w^2_{i\ell} \geq (L_2+1-n_i)\ \varepsilon^2.$$

We now apply the above concepts to the data for the two problem settings discussed earlier.

8.4. SOLUTIONS TO APPLICATIONS

8.4.1 R&D Project Efficiency Evaluation

When model (8.6') is applied to the data of Table 8-1, the efficiency scores obtained are as shown in Table 8-5.

Table 8-5. Efficiency Scores (Non-ranked Criteria)

Project	1	2	3	4	5	6	7	8	9	10
Score	0.76	0.73	1.00	0.67	1.00	0.82	0.67	0.67	0.55	0.37

Here, projects 3 and 5 turn out to be ‘efficient’, while all other projects are rated well below 100%. In this particular analysis, ε was chosen as 0.03. In another run (not shown here) where $\varepsilon = 0.01$ was used, projects 3, 5 and 6 received ratings of 1.00, while all others obtained somewhat higher scores than those shown in Table 8-5. When a very small value of ε (ε=0.001) was used, all except one of the projects was rated as efficient.

Clearly this example demonstrates the same degree of dependence on the choice of ε as is true in the standard DEA model. See Ali and Seiford (1993).

From the data in Table 8-1 it might appear that only project 3 should be efficient since 3 dominates project 5 in all factors except for input 5 where project 3 rates fourth while project 5 rates fifth. As is characteristic of the standard ratio DEA model, a single factor can produce such an outcome. In the present case this situation occurs because $w_{25}^2 = 0.03$ while $w_{24}^2 = 0.51$. Consequently, project 5 is accorded an 'efficient' status by permitting the gap between w_{24}^2 and w_{25}^2 to be (perhaps unfairly) very large. Actually, the set of multipliers which render project 5 efficient also constitute an optimal solution for project 3.

If we further constrain the model by implementing criteria importance conditions as defined in the previous section, the relative positioning of some projects change as shown in Table 8-6.

Table 8-6. Efficiency Scores (Ranked Criteria)

Project	1	2	3	4	5	6	7	8	9	10
Score	0.71	0.72	1.00	0.60	1.00	0.80	0.62	0.63	0.50	0.35

Hence, criteria importance restrictions can have an impact on the efficiency status of the projects.

8.4.2 Evaluation of Telephone Office Efficiency

The data of Table 8-4 has been evaluated using Model (8.6'). Both CRS and VRS models were applied, the results of which are presented in Table 8-7.

Initially, in applying DEA in this application, no attempt was made to impose constraints on multipliers. Under the CRS structure, approximately half of the offices (17 of the 33) are declared efficient. With the VRS model, the number of efficient units climbs to 25 out of 33. When criteria importance is introduced, the efficiency status (efficient versus inefficient) changes for some units. As well, the relative sizes of efficiency scores change. Note, for example, that the relative positions of offices 10 and 11 are reversed under the constrained VRS model versus those assumed in the unconstrained model. As well, only 15 of the offices (rather than 25) are rated as being efficient.

Table 8-7. Efficiency Scores

DMU#	CRS Score	VRS Score	VRS Score-constrained
1	1	1	1
2	1	1	1
3	1	1	1
4	.927	1	.973
5	1	1	.921
6	.907	.994	.906
7	.848	.849	.823
8	.668	.670	.644
9	.848	.970	.885
10	.617	.747	.731
11	.763	.815	.716
12	1	1	.915
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	.886
17	.898	1	1
18	.928	1	.935
19	.993	.993	.961
20	1	1	1
21	1	1	1
22	1	1	1
23	.846	1	1
24	.918	1	.904
25	1	1	1
26	1	1	.955
27	.824	.937	.926
28	.954	1	.919
29	.949	1	1
30	1	1	1
31	1	1	.907
32	1	1	1
33	.962	1	1

8.5. DISCUSSION

We have examined in this chapter the issue of performance measurement in the presence of qualitative data. The methodology presented herein demonstrates that when the idea of rank position data is introduced within the DEA structure, the resulting model can be transformed to a version of the conventional VRS model. This implies that all of the output results from standard DEA models apply. The CRS and VRS scores achieved using the model (8.6') are close to those obtained using the alternative IDEA structure of Cooper et al. (1999). This hints at the potential equivalence of the two approaches.

An important observation regarding radial projection, both here and in the IDEA approach of Cooper et al. (2001), is that one assumes that a $(1-\theta)$ x 100% reduction in a rank order position for an inefficient DMU, results in a legitimate (projected) rank order position. Of course, since radial projection treats all scales as continuous, not discrete, it would rarely be the case that projected points on the frontier would in fact correspond to discrete (Likert scale) positions. Hence, efficiency scores obtained by model (8.6') really represent lower bounds (on θ), and would in practice need to be adjusted upward to bring the projected positions to points that are allowable in Likert scale sense. We do not pursue herein how such adjustments would be made, but point to this as an interesting direction for future research.

REFERENCES

1. Banker, R.D., A. Charnes and W.W. Cooper (1984), Some models for technical and scale efficiencies in data envelopment analysis, *Management Science* 30, 1078-1092.
2. Cook, W.D., M. Hababou (2001), Sales performance measurement in bank branches, *OMEGA – International Journal in Management Science* 29, 299-307.
3. Cook, W.D., M. Hababou and H. Tuenter (2000), Multicomponent efficiency measurement and shared inputs in Data Envelopment Analysis: An application to sales and service performance in bank branches, *Journal of Productivity Analysis* 14, 209-224.
4. Cook, W.D., M. Kress and L. Seiford (1993), On the use of ordinal data in Data Envelopment Analysis, *Journal of the Operational Research Society* 44, 133-140.
5. Cook, W.D., M. Kress and L. Seiford (1996), Data envelopment Analysis in the presence of both quantitative and qualitative factors, *Journal of the Operational Research Society* 47, 945-953.
6. Cooper, W.W., K.S. Park and G. Yu (1999), IDEA and AR-IDEA: models for dealing with imprecise data in DEA, *Management Science* 45, 597-607.
7. Cooper, W.W., K.S. Park and G. Yu (2001), IDEA (Imprecise Data Envelopment Analysis) with CMDs (column maximum decision making units), *Journal of the Operational Research Society* 52, 176-181.
8. Kim, S.H., C.G. Park and K.S. Park (1999), An application of Data Envelopment Analysis in telephone offices evaluation with partial data, *Computers & Operations Research* 26, 59-72.

9. Thrall, R.M (1996), Duality, classification and slacks in DEA, *Annals of Operations Research* 66, 109-138.
10. Zhu, J. (2003), Imprecise Data Envelopment Analysis (IDEA): A review and improvement with an application, *European Journal of Operational Research* 144, 513-529.

This chapter is based upon Cook, W.D., M. Kress, and L. Seiford, On the use of ordinal data in DEA, Journal of the Operational Research Society, 44(2):133-140, and Cook, W.D., M. Kress, and L. Seiford, DEA in the presence of both quantitative and qualitative data, Journal of the Operational research Society, 47: 945-953, with permission from Palgrave.

Chapter 9

RESOURCE ALLOCATION IN AN R&D DEPARTMENT

9.1. INTRODUCTION

This chapter examines a model for aiding in resource allocation in a research and development setting. A particular organization investigated was the former Ontario Hydro. The models presented herein were designed using that setting as a framework; a framework that typifies a number of R&D situations. The basic problem involves allocating an annual research budget among a set of program areas and across various departments involved in those program areas, such that the overall benefit to the organization is as high as possible. While the problem of defining appropriate research areas is an important issue for management to consider, it will be assumed for purposes of the present discussion that these areas have already been decided.

The problem of resource allocation in regard to R & D departments has been approached primarily from the point of view of "projects" or "research options." Most of the models for assessing options assume multiple criteria or factors are involved. Utility approaches to the multidimensional problem have been examined by Geoffrion et al (1976), Hoel and Lin (1971) and Souder (1967 & 1973). Another class of approaches involves mathematical programming models such as those of Gear and Lockett (1972), and Charnes and Steadry (1964). Cook and Roll (1988) approach the problem of R & D project selection from the perspective of the past productivity of the organizations or operating units proposing the projects. Oral et al (1991) examine project prioritization from the point of view of data envelopment analysis.

In these approaches, it is assumed that data pertaining to each criterion appears in *quantitative* form. A few authors have concentrated on modelling in an environment in which only ordinal data can be obtained. See, for example, Bernardo (1977), and Cook and Seiford (1982). Lootsma et. al. (1986) present a method for deriving criteria weights based upon subjective judgments of executives, and then suggest allocating an available budget based on the best aggregate benefit to cost ratio. Here, the aggregate benefit for a project is the weighted total score obtained by that project on a set of ordinal factors.

All of these models assume that the problem is "zero-one," in the sense that a project or option is to be *funded or not funded.* They do not address the issue of partial allocation of resources to an initiative. Saaty (1980) suggests a method for doing this by basing the level of funding for an initiative on the rating that the initiative received in a weighted sense relative to the other initiatives being evaluated. Specifically, if on a 100 point scale one research option ranked at 85% while another ranked at 75%, Saaty's proposal calls for splitting the budget by directly using these two subjective ratings.

Other than this method suggested by Saaty (1980), no definitive resource allocation model in a qualitative data/multiple criteria environment presently exists. In the sections to follow, various issues pertaining to the resource allocation problem are discussed. A number of possible methods for allocating the research budget are presented, and a suggested approach is given.

9.2. CRITERIA FOR EVALUATING RESEARCH IMPACT

In a large research organization such as Ontario Hydro resource allocation must in general be viewed from a macro, i.e., program, point of view. Effects of shifts in budget allotments cannot normally be translated into impacts at a specific project level, but rather must be evaluated in terms of overall influences. Furthermore, impacts must be measured in terms of *various* factors or criteria. Specifically, various types of benefits are derived from the research in a given program (impacts on the environment, savings to the organization, ...). At the same time the capability of the department to carry out research, even if financial resources are not a constraint, may be influenced by the availability of certain liquid resources such as technical expertise, equipment, etc.

In Appendix A, a suggested set of criteria for evaluating the research program of a particular organization is given. A notable feature of some, if not all, of these factors is that they are generally measurable only on an ordinal scale. That is, if a research initiative is viewed in terms of a factor such as "dollars saved by the organization," it is very likely that the most accurate information one can expect a manager to supply regarding the impact that budget changes would have on this initiative, will be of the form "the effect is high, medium or low." Alternatively, one might adopt a 5-point rating scale where the levels are defined as

1 - the research has a *very high* impact in terms of this criterion
2 - has a *high* impact
3 - has a *medium* impact
4 - has a *low* impact
5 - has *almost no* impact.

Such a scale is common in many situations where qualitative factors must be evaluated.

9.3. INFORMATION REQUIREMENTS FROM MANAGEMENT

The resource allocation problem, as presented, has to do with finding the most advantageous way in which to partition the total research budget across a set of research areas. The problem must be viewed at this macro level of "research areas," rather than at the micro "specific project" level. Thus, the problem is *not* one of *project selection,* as will be discussed in the next chapter, but rather is one of "should \$X or \$Y be allocated to a given area?" In this context, the resource allocation problem must be viewed as a zero sum game – whatever additional budget any one research area receives, this will be at the expense of other areas.

Stated in simple terms, the resource allocation problem is:

"Given an existing research budget allocation to a set of research areas, can an improvement (in total benefit derived by the organization) come about by moving funds from one area to another?"

The answer to this question, in equally simplistic terms, is that this reallocation of research funds should be carried out if the *net loss* to the research area incurring the budget cut (and the loss to the organization at large) is less than the *net gain* to the area experiencing the budget increase. Clearly then, the key ingredient needed to be able to answer this question is information regarding the *marginal* net losses and gains associated with budget changes.

In the section to follow, a model is presented for helping management to evaluate the impact of such budget shuffles. This model requires, as input data, managements' opinions as to the *marginal effects*, in terms of each criterion, that are likely to be created by budget modifications. The simplest way to view budget modifications is in terms of some *basic monetary unit.* While the size of this basic unit must ultimately be decided by management, it is assumed for illustrative purposes herein that a basic unit of $100,000 is to be used.

As a starting point, suppose that management supplies information to the following questions:

1. For each research area how would you rate, on a five point scale, the impact of a budget cut of one basic monetary unit? Make this judgment for each of the criteria (Appendix A).
2. What is the effect of a $2 unit cut?
3. What is the effect of a $1 unit increase?
4. What is the effect of a $2 unit increase?

It must be noted at this point that the above questions regarding a $2 increase or decrease can be posed in at least two ways. The first is that given above: "What is the effect of a $2 unit change?" Alternatively, the question could be presented as: "What is the effect of a second $1 unit change *given* that a $1 unit change has already taken place?" While at present it is not clear why we distinguish between these two information solicitation modes, this will become more apparent later. For purposes of the development in this and the next section, we assume the questions as originally posed are appropriate.

Table 9-1. Response Matrix

Research Area j	Actions Change in (Budget/Level)	Criterion $\beta 1$	$\beta 2$	...	$\beta 7$
1	$\delta 1 = -\$2$ unit	1	2	...	1
	$\delta 2 = -\$1$ unit	2	3	...	3
	$\delta 3 = +\$1$ unit	2	1	...	2
	$\delta 4 = +\$2$ unit	1	1	...	3
2	$\delta 1 = -\$2$ unit	2	4	...	1
	$\delta 2 = -\$1$ unit	4	1	...	3
	$\delta 3 = +\$1$ unit	3	4	...	1
	$\delta 4 = +\$2$ unit	2	4	...	5

Responses to these questions can be recorded in the form of a matrix for each research area. The entries inside this *response matrix* are numbers from the 5-point scale discussed earlier. Table 9-1 illustrates such responses. The "1" in the upper left hand corner of the matrix, for example, would mean that a \$2 unit budget cut to research area p is rated as having a *very high detrimental* effect on that research area in regard to benefit β_1.

The process whereby ratings such as these can be derived is an important management issue. It is advisable to create the table in a "column- wise" fashion. Thus, for each criterion, the decision maker would first need to establish the impact, on a 5-point scale, that a \$1 unit budget increase or decrease would entail. Using this as a *bench mark*, a \$2 unit change in either direction would then be evaluated, and so on. This approach of working in a column allows for comparison of a budget change of one size to a budget change of another size. Evaluating a single budget change across all criteria first (i.e., "row wise") is a much more difficult task.

It must be emphasized that the particular set of criteria to be used is not an issue here (these must be decided by management), nor is the number of possible actions (-\$2, -\$1, +\$1, +\$2). The latter can be expanded to any number of incremental budget changes desired. Furthermore, it may be desirable to think in terms of % changes to research area budgets, rather than specific \$ amounts. This matter will be discussed later. What is at issue here is the need to think in terms of *marginal impacts* if budget modifications are made. It is information of this type which will drive the model discussed in the next sections.

9.4. MODELLING RESOURCE ALLOCATION – THE BASIC IDEA

The basic approach recommended for allocating the research budget among a set of research areas consists of four steps:

Step 1: Obtain rating information of the type presented in Table 9-1 for each research area;

Step 2: Determine, using a multiple criteria model (various models are described in the next subsection), a rating R_p^v for each research area p and budget change v. This rating will be a percentage $\leq 100\%$, and will represent the relative impact of that budget change on that area. This impact can be positive (if $v > 0$) or negative (if $v < 0$);

Step 3: Arrange the R_p^v values from lowest to highest;

Step 4: Make budget reallocations according to the relative sizes of the R_p^v.

The mechanics of Steps 2 and 4 will be described in the following sections. Various techniques are proposed there for using the qualitative data of Step 1, and arriving at a set of weights $w_{k\ell}(p,v)$. These weights represent the worth or value of being ranked in the ℓ th rank position, on the 1 to 5 scale, relative to criterion k. The $w_{k\ell}(p,v)$ [1] for each research area p and budget level change v are combined to provide an overall rating R_p^v.

To illustrate Steps 3 and 4, suppose that there are only 3 research areas, and let the R_p^v values be as given in the following table.

Table 9-2. R_p^v Values

Research	Budget Change v			
Area p	- \$2	- \$1	+ \$1	+ \$2
1	.83	.76	.61	.88
2	.79	.72	.70	.72
3	.65	.81	.56	.58

Arranging these values from highest to lowest we have:

$$R_1^{+2} = .88$$
$$R_1^{-2} = .83$$
$$R_1^{+1} = .81$$
$$R_2^{-2} = .79$$
$$R_2^{-1} = R_2^{+2} = .72$$
$$R_2^{+1} = .70$$
$$R_3^{-2} = .65$$
$$R_3^{-1} = .61$$
$$R_3^{+2} = .58$$
$$R_3^{+1} = .56$$

The idea is that we can shift an amount of funds $v = \$\ell$ from any research area p_1 to an area p_2 provided the corresponding $R_{p_1}^{-\$\ell} < R_{p_2}^{+\$\ell}$. For example, the negative benefit or damage associated with a \$2 unit decrease in the budget of area 3 is rated at $R_3^{-2} = .65$. On the other hand, a \$2 unit increase in the budget of area 1 is rated at $R_1^{+2} = .88$. Thus, a shift of \$2 from area 3 to area 1 produces a net gain in benefits to the organization. We define the *gain index* for the movement of \$2 units between research areas 3 and 1 as

[1] We use the notation $w_{k\ell}$ only from this point on.

$$G_{31}^{2} = f(R_3^{-2}, R_1^{+2}).$$

Clearly, any movement of funds that results in a positive gain index will improve the overall benefit of the research budget to the organization.

The problem to be solved, therefore, is to find a set of gain indices $G_{p_1p_2}^{v}$ which represent the best improvement in overall research impact to the organization. In a later section we present a possible function f for defining the $G_{p_1p_2}^{v}$, and a heuristic method for deriving an appropriate optimal set of $G_{p_1p_2}^{v}$ values is discussed.

It is pointed out that Steps 1-4 represent a reallocation of the budget relative to a *current budget split.* Once this reallocation has been completed, management may decide to repeat the entire exercise, beginning with the re-specification of perceived marginal changes in benefits relative to the newly created budget split. As many cycles could be carried out as felt necessary by management.

9.5. DEVELOPING THE RATINGS R_p^v

The R_p^v are to represent a prioritization of the overall benefits associated with various budget changes that could be applied to the set of research areas under consideration. For purposes of the discussion in this section, we will cast the problem in a somewhat more generic setting than that of the previous sections. That is, suppose there are N items or alternatives (budget changes, projects, products, etc.) to be ranked. Each of these is to be evaluated in terms of K ordinal or qualitative criteria. In the example described earlier (Section 4), each (budget change, research area) pair constitutes an alternative. Thus, in that setting N was equal to $3 \times 4 = 12$.

For illustrative purposes, consider the situation in which six alternatives, e.g., budget changes, applied to research areas are to be evaluated in terms of three criteria - (1) safety benefits, (2) environmental benefits, and (3) long term returns (profitability). Suppose that for each criterion the six alternatives can only be evaluated in an *ordinal* sense, say on a 5-point scale.

In the case of safety, for example, alternative 1 ranks last or 5th, alternative 2 ranks 3rd, ... etc. Table 9-3 displays the evaluations for all alternatives on all criteria.

The problem at hand is to prioritize or rank the six alternatives from most to least preferred, utilizing the *preference* data provided. At least two complicating issues must be addressed, however, if such a prioritization is to be achieved. *First,* the standing or importance of a given alternative may be very different on some criteria than on others. Alternative #1, for example, ranks 1st on criterion 2, but last (5th) on criterion 1. On the other hand, alternative #3 ranks 1st on criterion 1 but 3rd on criterion 2. Which alternative should be ranked highest? This question raises a *second* issue of *criteria importance*. If criterion 1 is much more important than criterion 2, how should this be factored into the analysis? In some cases it may be possible to supply reasonably representative weights, while in other instances it may only be possible to *rank the criteria*. How should an ordinal ranking of criteria be dealt with?

Table 9-3. Rankings of Alternatives by Criteria

	Criteria		
Alternative	1	2	3
1	5	1	2
2	3	4	5
3	1	3	4
4	2	3	4
5	2	2	2
6	4	5	1

9.5.1 Conventional Approach

The problem of how to combine multiple attribute data is a familiar problem of utility theory. Essentially, we want to develop an overall utility function which can reduce the problem to one involving a *common quantitative* unit of measurement.

A crude, but often utilized procedure for obtaining a prioritization of alternatives in such situations is to begin with a given set of criteria weights w_1, w_2, w_3. Using these "known" values, a *weighted rank* is obtained for each alternative. These weighted ranks are then arranged from lowest to highest to achieve the desired prioritization of the alternatives. Suppose, for example, that safety is given an importance weight of $w_1 = 10$, environment a weight $w_2 = 7$ and profitability a weight $w_3 = 5$. The weighted rank R_1 for alternative #1 would then be

$$R_1 = 10 \times 5 + 7 \times 1 + 5 \times 2 = 67.$$

The corresponding value R_2 for alternative #2 is

$$R_2 = 10 \times 3 + 7 \times 4 + 5 \times 5 = 83.$$

For this set of weights, alternative #1 comes out at a lower value than #2, meaning that project #1 should be given a higher priority than #2 [2].

[2] Since a rank of 1 means "most important," and 5 means "least important," the project with the lowest weighted rank will be given the *highest* priority. Clearly, if we reversed the scale (5 is best, 1 is worst), the opposite interpretations would be given to R_1 and R_2.

There are two basic operational shortcomings with this crude approach:

1. The method requires that the analyst be able to choose, in some manner and using some scale, a set of weights reflecting the *absolute* importance of the criteria. While in some instances criteria weights may have evolved over time and are a "given", in other cases the weight assignment exercise is ad hoc, hence very much at the *whim* of the decision maker. Even when such assignments are based on the very best advice and information from the relevant players at the time, the scales and values chosen are, in the final analysis, arbitrary. Furthermore, the values chosen often arise from a set of widely varying opinions solicited from experts, executives, etc. The final "consensus" may be less than satisfactory.

2. A second, and even more disturbing aspect of this methodology is the fact that the rank positions of the alternatives, which are only intended as *relative* (ordinal scale) priorities, are being treated as if they were *absolute* cardinal (interval scale) values. Ranking alternative 3 in 1st place on the safety criterion, and alternative 4 in 2nd place, for example, is *not* meant to imply that alternative 3 should be valued as being *twice as important* as alternative 4 relative to this criterion. These rank positions express relative priorities only, not absolute worths.

9.5.2 A Proper Evaluation of Ordinal Data

Therefore, there are generally two sets of "unknowns" in such an environment — the v_ℓ, expressing the "value" of the different rank positions $\ell = 1,\dots,L$, and the w_k expressing the importance of the K criteria.

With this notation, if an alternative ranks in the ℓ^{th} category or position on the k^{th} criterion, it will be given a credit of $w_k v_\ell$ for this criterion. Recall that in the above example, alternative #1 received a total credit or value of

$$R_1 = w_1 \times 5 + w_2 \times 1 + w_3 \times 2.$$

Thus, for criterion 1, the credit was $w_1 \times 5$. The suggestion is that the credit should be $w_1 v_5$, where v_5 and w_1 are to be determined. As a general setup, let us use a single variable, with a double subscript, $w_{k\ell}$ in place of the product $w_k v_\ell$. In this case the restrictions specified in (a) and (b) above become

$$w_{k\ell} > w_{k\ell+1} \tag{9.1}$$

$$w_{k\ell} > w_{k+1\ell}, \tag{9.2}$$

for all ℓ and k.

Clearly, there are an infinite number of combinations of weights $w_{k\ell}$ satisfying these conditions. What is required is a procedure for selecting an "appropriate" set. Whatever this set is, it will immediately dictate the "rating" which each alternative will receive, and therefore the final rank ordering of the alternatives. For example, in the case of alternative #1 in the above, R_1 is given by

$$R_1 = w_{15} + w_{21} + w_{32}, \tag{9.3}$$

since this alternative ranked fifth on criterion 1, first on criterion 2 and second on criterion 3.

One approach for deriving a set of worths $w_{k\ell}$ is to use a philosophy similar to the Data Envelopment Analysis (DEA) method as proposed by Charnes et al (1978). This approach strives to find for each of a set of alternatives (budget shifts) a best or highest possible rating subject to certain constraints on the weights $w_{k\ell}$. In the present case, this would amount to maximizing R_p^v for each area p and budget shift v. Proceeding in this fashion, a set of multipliers $w_{k\ell}$ would be determined corresponding to each (p, v) pair. Cook and Kress (1991) present a modified version of this model wherein a single set of multipliers can be derived. We adopt this latter approach here. Stated in basic and general terms, the model for deriving the $w_{k\ell}$ takes the form

$$\max\ z \tag{9.4a}$$

subject to

$$R_p^v \le 100\%, \text{ for all } p, v \tag{9.4b}$$

$$w_{kl} - w_{k+1\ell} - g_{k\ell}(z) \ge 0, \quad \text{for all } \ell \quad k = 1, \ldots, K-1 \tag{9.4c}$$

$$w_{K\ell} - g_{K\ell}(z) \ \ge 0, \quad \text{for all } \ell \tag{9.4d}$$

$$w_{k\ell} - w_{k\ell+1} - h_{k\ell}(z) \ \ge 0, \quad \text{for all } k, \quad \ell = 1, \ldots L-1 \tag{9.4e}$$

$$w_{kL} - h_{kL}(z) \ \ge 0, \quad \text{for all } \ k, \tag{9.4f}$$

where k is the set of criteria under consideration, L is the number of rank positions for evaluating impacts (L=5 in the present example), and $g_{k\ell}(z)$ and $h_{k\ell}(z)$ are discriminating functions (see Cook & Kress (1991)).

The constraints (9.4b) restrict the aggregate rating of each budget shift v per program p to not exceed 100%. Constraints (9.4c) and (9.4d) specify that the extent to which one discriminates between criteria of consecutive importance (k and $k+1$) should be at least some amount $g_{k\ell}(z)$. Constraints (9.4e) and (9.4f) specify this same discrimination vis-a-vis rank position ℓ and $\ell+1$. Finally, the objective of maximizing the discrimination parameter z is intended to uncover a set of weights $w_{k\ell}$ that provide some form of maximum discrimination between consecutive rank positions and between criteria of decreasing importance k.

The theoretical properties and rationale for such a model is provided in Cook and Kress (1991).

Later in the chapter a full example is given illustrating the output from this model.

9.5.3 Multiple Voters and Levels of Credibility

It is useful to view the structure of the research division in terms of research areas and departments. Viewed in matrix format, each department is involved in research in one or more of the areas. Thus, we can view the matrix as consisting of 0s and 1s. There is a 1 in the (d, p) slot of the matrix if department d allocates part of its budget to work on area p. Otherwise, there is a 0 in the slot.

	Research Area p				
Dept. d	1	2	3	⋯	10
1	1	0	1	⋯	0
2	0	1	1	⋯	0
3	0	0	1	⋯	1
⋮					
6					

In the process of gathering management opinions as to the impacts of budget changes, it must be assumed that each "voter," say a department manager, will provide a set of information as described in the previous sections. The problem arises as to how to *aggregate* the opinions of say, six department managers. At least two complicating factors must be considered. First, for any given research area p only a subset V_p of the voters may provide an opinion $(V_p \leq 6)$. Second, because a given voter may have more knowledge of a research area than some other voter may have, it is necessary to consider the "credibility" of the opinions expressed.

To illustrate the ideas to be discussed, consider the simple example where there are 2 departments, 2 research areas, 1 criterion and 2 proposed budget changes. Furthermore, suppose that any manager's competence or credibility is graded at one of three possible levels:

Level 1 - highly competent
Level 2 - competent
Level 3 - low competence
Let the *credibility matrix* be

	Research Area	
Dept.	1	2
1	1	3
2	2	1

Thus, department manager #1 has full knowledge of area 1 (credibility level 1), but has poor knowledge of area #2 (credibility level 3). Suppose the responses for the two managers are:

Manager #1

Program	Δ Budget	r_{pck}^{d} Criterion 1
1	Δ_1	2
	Δ_2	1
2	Δ_1	3
	Δ_2	2

Manager #2

Program	Δ Budget	r_{pck}^{d} Criterion 1
1	Δ_1	1
	Δ_2	1
2	Δ_1	2
	Δ_2	2

That is, manager #1 believes that a budget change Δ_1 (e.g. a $100,000 increase) to research area 1 would have an impact relative to criterion 1 equal to "2" on a 5-point scale, and so on. Manager #2, however, rates the impact as a "1," (that is, he/she believes it to have more impact.)

The problem is to compile or aggregate the opinions of the managers into one *overall* set of responses.

The approach suggested determines *weights* to be used in combining managers opinions. These weights are then applied to the individual ratings to get a *weighted median.* We introduce the following notation:

p = index for the research area under consideration
c = index for the criterion under consideration
j = index for the budget change under consideration
d = index for the department under consideration
V_p = a constant representing the number of managers who voted on research area p
L_p = set of competence levels of managers

who vote on research area p.

r^d_{pcv} = the rating which department manager d gives to budget change v concerning criterion c and research area p

W_ℓ = a decision variable describing the worth of an opinion with ℓ^{th} ranked competence level $(\ell = 1, 2, 3)$ that a department manager gives regarding a research area.

Using arguments similar to those of the previous sections we can derive an appropriate set of multipliers W_1, W_2, W_3 by solving the following pre-emptive linear programming problem.

$$\max \quad Mz + \sum_p \sum_c \sum_{v \varepsilon V} \sum_d W_{\ell_{pd}} \, | r^d_{pcv} - r_{pcv} |$$

subject to

$$\begin{aligned} W_1 - W_2 - g_1(z) &\geq 0 \\ W_2 - W_3 - g_2(z) &\geq 0 \\ W_1 + W_2 + W_3 &= C \end{aligned} \tag{9.5}$$

where M is a large constant, V is the set of all budget changes (note some members of V are negative changes and some are positive), $g_\ell(.)$ is a discriminating function, and C is some scaling constant. The variables in this LP problem are r_{pcv}, W_ℓ, and z.

Here, the notation ℓ_{pd} denotes the credibility level for department manager d when voting on program p. For example, referring to the credibility matrix above, $\ell_{21} = 3$. That is, manager #1 is not very familiar with program #2 and is rated as having low competence or reliability in terms of his/her vote. Hence, the rating level is 3.

Suppose that $g_\ell(z) = z$ for all ℓ, then it can be shown that the solution to this problem is such that $W_3^* = z^*, W_2^* = 2z^*, W_1^* = 3z^*$. That is, the W_i^* are a type of Kendall score (Kendall (1962)). This being the case, it is not necessary to solve the above problem to find the appropriate median ratings $r_{pcv.}$ If, for example, $C = 6$, then $W_1^* = 3, W_2^* = 2, W_1^* = 1$, (Kendall scores) and the "weighted" median of the $\{r^d_{pcv}\}_d$ will yield the required r_{pcv}.

This approach then permits us to aggregate managers' opinions into an overall consensus set of ratings.

In this section the problem of evaluating the relative impacts of various budget modifications at the program level has been examined. We now look at the issue of deriving a best *set* of budget adjustments.

9.6. BUILDING REALLOCATION

Earlier the idea of a gain index $G^v_{i_1i_2}$ was presented, specifically

$$G^v_{i_1i_2} = f(R^{-v}_{i_1}, R^v_{i_2}).$$

Since the R^v_i are relative *ratings* rather than numerical quantities (such as monetary impacts) there is no clear definition for the function f. Since the R^v_i arise from a linear model (hence, are a form of linear or additive utility) a reasonable definition for f is:

$$G^v_{i_1i_2} = f(R^{-v}_{i_1}, R^v_{i_2}) = R^v_{i_2} - R^{-v}_{i_1} \qquad (9.6)$$

That is, f is the net gain or improvement in the rating by moving resources v from area i_1 to area i_2. This idea is used below.

At this point it is necessary to distinguish two cases pertaining to the manner in which monetary units shift between areas:

Case I: Monetary Shifts with no Splitting

In this case it is assumed that if v monetary units are moved *from* area p_1, that same v unit is moved *to* one and only one other area p_2 (the v units cannot be split across a *set* of areas). The following zero-one integer programming problem can be used to determine an optimal set of budget shifts:

$$\max \quad \sum_{p1}\sum_{p2}\sum_{v} G^v_{p1p2} x^v_{p1p2} \qquad (9.7a)$$

subject to

$$\sum_{p2\neq p1}\sum_{v} x^v_{p1i2} \leq 1 \quad \text{for all } p_1 \qquad (9.7b)$$

$$\sum_{p1\neq p2}\sum_{v} x^v_{p1p2} \leq 1 \quad \text{for all } p_2 \qquad (9.7c)$$

$$\sum_{p2\neq p}\sum_{v} x^v_{pp2} + \sum_{p1\neq p}\sum_{v} x^v_{p1p} \leq 1, \quad \text{for all } p \qquad (9.7d)$$

$$x^v_{p1p2} = 0 \text{ or } 1 \qquad (9.7e)$$

Here,

$$x^v_{p1p2} = \begin{cases} 1 \text{ if } v \text{ monetary units are moved from area } p_1 \text{ to area } p_2 \\ 0 \text{ otherwise} \end{cases}$$

Constraint (9.7b) ensures that at most one level of resources (v) leaves any area p_1 and requires that if this happens then this amount go to exactly one other area p_2. Constraint (9.7c) guarantees that any amount v entering an area p_2 can come from exactly one source p_1. Constraint (9.7d) prohibits resources from both entering and leaving any given research area p.

This integer programming problem can be viewed as a generalization of a multilevel linear assignment problem with the set of matrices $\{G^v_{p_1p_2}\}_{v\varepsilon M}$ constituting the levels. The generalization arises from the fact that what will constitute the origins and destinations will come about as a result of the optimization process rather than being known beforehand. That is, if resources *leave* an area p_1, and go to an area p_2, then p_1 will, by definition, be an origin and p_2 a destination. If one knew in advance which areas were origins, which destinations and which neither, then constraint (9.7d) could be ignored and (9.7a),(9.7b),(9.7c),(9.7d) would be solved as a standard linear assignment problem.

The situation is further complicated by at least two possibilities which may arise:

1. It may be beneficial to move resources in *either* direction between two areas. Specifically, there can exist pairs of areas (p_1, p_2) for which $G^v_{p_1p_2} > 0$ *and* $G^v_{p_2p_1} > 0$ as well. This can be the case, for example, if two areas are such that gaining resources in each is highly beneficial, while losses in resources from either has little impact;
2. A program p_1 can be an *origin* relative to some areas p_2 (it is beneficial to move resources from p_1 to p_2), but a *destination* relative to other areas p_2.

In the particular problem under investigation, a conventional zero-one integer programming algorithm was used.

Case II: Monetary Shifts with Splitting

In this case it is assumed that when v_1 monetary units leave p_1, these v_1 units can be split among several other areas p_2. Similarly, the v_2 units that enter any area p_2 can come from several areas p_1.

In this situation, the issue arises as to how to measure the *net loss (gain) per unit* of resources leaving (entering). If, for example, $v_1 = 3$ monetary units, rated as $R^3_{p_1}$ in terms of loss to area p_1, and if 1 of those units is moved to area p_2 and the other 2 units to p_2, how should the gains and losses be evaluated? If we were to argue that the loss per unit to area p, is $R^3_{p_1}/3$ (or in general $R^v_{p_1}/v$), there arises a scaling problem in the sense that the maximum value of $R^2_{p_1}$ is 100%. So, if v is large then the loss per unit to area p, is very small (hence unimportant), and it will always appear beneficial to move a large amount of resources out of an area and split that amount among as many other areas as possible. There is an additivity (or divisibility) problem here.

There is no clear resolution to this problem, and we therefore do not presume to be able to provide the definitive answer. We do offer, however, one possible approach below. In describing this approach, however, it is necessary to return to the issue of how information is solicited from management. Recall that the point was made earlier that the methods used to elicit information from management can take different forms. Clearly, we get different outcomes in terms of the R_p^v depending upon the approach used to gather the information.

If *splitting* is permitted, it is reasonable to use the *second* form of information elicitation mentioned earlier. That is if v=\$3, for example, then three different ratings $R_{p_1}^1, R_{p_1}^2, R_{p_1}^3$ are obtained using managements responses to the questions

- What is the impact of the 1st \$1 change?
- What is the impact of the 2nd \$1 change?
- What is the impact of the 3rd \$ change?

Furthermore, define $\overline{R}_{p_1}^v = \sum_{\ell=1}^{v} R_{p_1}^\ell$. So

$$\overline{R}_{p_1}^1 = R_{p_1}^1, \overline{R}_{p_1}^2 = R_{p_1}^1 + R_{p_1}^2, \overline{R}_{p_1}^3 = R_{p_1}^1 + R_{p_1}^2 + R_{p_1}^3.$$

Hence, $\overline{R}_{p_1}^v$ is the "total" impact of a $\$v$ resource shift. Using this definition, the implied average impact *per monetary unit* is $\overline{R}_{p_1}^v/v$. This average value is used below in the model for deriving optimal resource shifts.

With this in mind, we generalize the definition of the $G_{p_1p_2}^v$ to

$$G_{p_1p_2}^{v_1v_2} = \frac{-\overline{R}_{p_1}^{v_1}}{v_1} + \frac{\overline{R}_{p_2}^{v_2}}{v_2}. \tag{9.8}$$

$G_{p_1p_2}^{v_1v_2}$ is then the average *net benefit* of *each* monetary unit moving from area p_1 to area p_2 if p_1 loses v_1 units in total and p_2 gains v_2 units in total. Define the integer variables $x_{p_1i_2}^{v_1v_2}$ = number of monetary units flowing directly from area p_1 to area p_2 where v_1 leaves p_1 and v_2 enters p_2. Define also the variables

$$y_{p_1}^{v_1} = \begin{cases} 1 & \text{if } v_1 \text{ monetary units leave } p_1 \\ 0 & \text{otherwise} \end{cases}$$

$$z_{p_2}^{v_2} = \begin{cases} 1 & \text{if } v_2 \text{ monetary units enter } p_2 \\ 0 & \text{otherwise} \end{cases}$$

An optimal set of budget shifts can then be derived from the following linear integer programming model:

$$\max \sum_{p_1}\sum_{p_2}\sum_{v_1}\sum_{v_2} G_{p_1 i_2}^{v_1 v_2}\, x_{p_1 i_2}^{v_1 v_2} \tag{9.9a}$$

subject to

$$\sum_{p_2}\sum_{v_2} x_{p_1 p_2}^{v_1 v_2} = v_1 y_{p_1}^{v_1} \quad \text{for all } p_1, v_1 \tag{9.9b}$$

$$\sum_{p_1}\sum_{v_1} x_{p_1 i_2}^{v_1 v_2} = v_2 z_{p_2}^{v_2} \quad \text{for all } p_2, v_2 \tag{9.9c}$$

$$\sum_{v_1} y_{p_1}^{v_1} + \sum_{v_2} z_{p_1}^{v_2} \geq 1 \quad \text{for all } p_1 \tag{9.9d}$$

$$\sum_{p_1}\sum_{v_1} v_1 y_{p_1}^{v_1} - \sum_{p_2}\sum_{v_2} v_2 z_{p_2}^{v_2} = 0 \tag{9.9e}$$

$$x_{p_1 p_2}^{v_1 v_2} \geq 0 \text{ and integer}$$

$$y_{p_1}^{v_1},\ z_{p_2}^{v_2}\ \varepsilon\ \{0,1\}$$

Constraint (9.9d) ensures that resources can either flow into or out of any given area (but not both) and in only one amount v. Hence, for any p_1 at most one variable in the set $\{y_{p_1}^{v}, z_{p_1}^{v}\}_v$ is positive. Given this, constraints (9.9b) and (9.9c) balance the flows in and out of areas. Finally, constraint (9.9e) guarantees the equality of total resources leaving all areas with total resources entering all areas.

9.7. APPLICATION

A particular organization examined involved 9 departments with 5 program areas. Total existing budgets for those areas are:

p_1 - \$3.5 million
p_2 - 7.3 million
p_3 - 4.6 million
p_4 - 9.3 million
p_5 - 8.2 million

Information was gathered from the 9 department heads involving:

(a) the ratings of the impacts of budget changes vis-a-vis four criteria. In this particular case a three point rating system was used (as opposed to a 5-point scale). The data provided shows the perceived impacts of *decreasing* a budget by \$0.5 million, \$1 million or \$2 million, and increases of these same amounts. It is noted that this data was collected with the intention of applying the restricted model (9.7a)-(9.7e).

(b) prioritization of the 4 goals.

Table 9-4 presents the ratings of the various goals by the 9 voters. A consensus model (Cook and Seiford (1978)) was used to aggregate these into a single ranking (Goal 1 > Goal 2 > Goal 4 > Goal 3). In Table 9-5 the *aggregated* or median of the 9 voters opinions as to the impacts of budget shifts is given. Applying model (9.4a)-(9.4f) the rating of the various budget shifts for the different programs were obtained, and appear in Table 9-6. In the application of the restricted model (9.7a)-(9.7e) (no splitting of monetary shifts) there are three alternative optima as shown in Table 9-7.

Table 9-4. Ratings of Goals by Voters (5-point Scale)

	Goal			
Voter	G1	G2	G3	G4
a	1	1	2	4
b	2	3	1	5
c	1	2	1	1
d	1	2	5	2
e	3	3	4	2
f	2	2	3	5
g	3	1	5	4
h	1	2	4	1
i	1	3	3	3

Table 9-5. Budget Change Impacts

	Budget	Goal			
Program	change	1	2	3	4
p1	+2	—	—	—	—
($3 million)	+1	3	2	2	3
	+.5	3	2	3	3
	-.5	3	2	3	2
	-1	3	1	2	1
	-2	3	1	2	1
P2	+2	—	—	—	—
($6.6 million)	+1	3	2	3	2
	+1	3	2	3	2
	+.5	3	3	3	3
	-.5	3	3	3	2
	-1	2	2	3	2
	-2	1	1	3	1
P3	+2	—	—	—	—
($5.5 million)	+1	2	2	1	2
	+.5	3	3	2	3
	-.5	3	3	2	2
	-1	3	2	2	1
	-2	3	1	2	1
P4	+2	—	—	—	—
($5.4 million)	+1	2	3	1	2
	+.5	3	3	2	2
	-.5	3	3	2	2
	-1	2	3	1	1
	-2	1	3	1	1
P5	+2	—	—	—	—
($10.4 million)	+1	3	2	2	3
	+.5	3	3	2	3
	-.5	3	3	2	3
	-1	3	3	3	2
	-2	3	2	3	2

An additional set of data was collected to be used for the general model (9.9a)-(9.9e). We do not bother to present the detailed data here, but do point out that the optimal budget shifts arising from this (as shown in Table 9-8) are quite different from those of the restricted model. It is noted that complete matching does not need to occur in this model. For example, two programs each lost $2, while three programs gain, with some gains being in smaller amounts (in two of the three cases).

Note that an optimal ranking of the goals when these 9 opinions are combined is Goal 1 > Goal 2 > Goal 4 > Goal 3. That is, the goals G1, G2, G3, G4 are ranked 1,2,4,3 respectively. A consensus ranking method was used to derive this overall ranking.

Table 9-6. Aggregate Ratings for Program Budget Combination

Alternative		Rating	
P2B6	100.0	-2-2	62.57
-3-1	92.8 *	-5-6	62.57*
-4-6	92.8	-1-2	61.13
-4-1	87.04**	-1-1	62.13
-3-2	84.17	-5-2	61.13
-2-1	82.73	-3-4	58.25
-1-6	81.29**	-4-4	58.25
-3-6	81.29	-4-3	58.25
-1-5	81.29	-1-3	56.81
-4-5	81.29	-2-4	53.94
-4-2	75.53	-5-5	53.94
-2-5	74.09	-5-4	52.50
-3-5	72.65	-5-4	52.50
-5-1	69.77	-3-3	52.50
-1-4	62.57	-2-3	48.18

Notation:
P2B6 - Program #2 & Budget change #6 (i.e. - $2)
P3B2 - Program #3 & Budget Change #2 (i.e., +$1)
Note:
* $2 moving from P5 to P3
** $2 from P1 to P4.

Table 9-7. Alternate Optimal Solutions

Scenario #	Solution 1	Solution 2	Solution 3
1	$1 from P5 ⇒ P3 $0.5 from P2 ⇒ P4	—	—
2	$2 from P5 ⇒ P3 $2 from P1 ⇒ P4	$1 from P5 ⇒ P3 $2 from P1 ⇒ P4	$2 from P1 ⇒ P3 $2 from P5 ⇒ P4
3	$2 from P5 ⇒ P4 $1 from P2 ⇒ P3	$2 from P5 ⇒ P3 $0.5 from P2 ⇒ P4	$1 from P5 ⇒ P3 $0.5 from P2 ⇒ P4

Table 9-8. Optimal Solution – Splitting Allowed

Program #	$ leaving	$ entering	Final Budgets
1	—	$0.5	$4
2	$	—	5.3
3	—	$1.5	$6.1
4	—	2.0	11.3
5	$2	—	6.2

9.8. CONLUDING COMMENTS

We have presented a model for evaluating budget shifts among a set of programs or research areas, where the impact data are ordinal. Such data is typical of this environment insofar as impacts on broad general research initiatives are difficult to quantify. In the process of making budget shifts among research areas, there are implied impacts on the sizes of the budgets held by the departments. If, for example, a department derives its entire budget as a result of research carried out in one area, then losses or gains in that area may have immediate severe consequences vis-a-vis that funding. On the other hand, if a department carries out research in a number of areas, some of which are down graded (budgets reduced) while others are upgraded, there may be no effect on that department's funding at all.

It may be possible to minimize budget change impacts at the department level (either by budget increases that can cause staff shortages, or decreases that may lead to staff layoffs) using a goal programming approach. Possible goals may be (a) to retain department budgets at current levels, (b) avoid layoffs in departments where staff may need to interact with other departments, (c) avoid increased staff needs of a type that is difficult to acquire, and so on. While these department-level impacts have not been addressed here, they are by no means trivial considerations. They are, however, a second level issue worthy of later study.

REFERENCES

1. Bernardo, J. (1977), An assignment approach to choosing R&D experiments, *Decision Science* 8, 489-501.
2. Charnes, A, W.W. Cooper, and E. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2, 427-444.
3. Charnes, A. and A. Stedrey (1964), Chance constrained model for real-time control in research and development management, *Management Science* 12, 353-362.

4. Chung, A., W. Cook, and M. Kress (1998), Resource allocation in R&D departments, *INFOR* 36, 41-57.
5. Cook, W. and M. Kress (1991), A multiple criteria decision model with ordinal preference data, *European Journal of Operational Research* 54, 191-198.
6. Cook, W. and Y. Roll (1988), R&D Project selection: productivity considerations, *R&D Management* 18, 251-256.
7. Cook, W. and L. Seiford (1982), R&D project selection in a multidimensional environment: A practical approach, *Journal of the Operational Research Society* 33, 397-405.
8. Cook, W. and L. Seiford (1978), Priority ranking and consensus formation, *Management Science* 24, 1721-1732.
9. Gear, A. and A. Lockett (1972), Programme selection in research and development, *Management Science* 18, B575-B590.
10. Geoffrion, A., L. Dyer, and A. Feinberg (1973), An interactive approach for multicriterion optimization with an application to the operation of an academic department, *Management Science* 19, 321-335.
11. Hoel, L. and F. Lin (1977), Structure of utility-based evaluation approaches, *Transportation Enqineering Journal* 9, 307-321.
12. Lootsma, F., J. Meisner, and F. Schellemans (1986), Multi-criteria decision analysis as an aid to the strategic planning of energy R&D, *European Journal of Operational Research* 25, 216-234.
13. Oral, M., O. Kettani, and P. Lang (1991), A methodology for collective evaluation and selection of industrial R&D projects, *Management Science* 37, 871-885.
14. Saaty, T. (1980), The Analytic Hierarchy Process, McGraw-Hill, New York.
15. Souder, W.E. (1967), Selecting and staffing R&D projects via operations research, *Chemical Engineering News* 63, 27-37.
16. Souder, W.E. (1973), Utility and perceived acceptability of R&D project selection models, *Management Science* 19, 1384-1394.

This chapter is based upon Chung, A, W.D. Cook, and M. Kress, 1998, Resource allocation in R&D departments, *INFOR*, 36, ½: 41-57, with permission from University of Toronto Press.

APPENDIX A: BENEFITS

Criteria	Sub-criteria or Interpretation
1.Enhancement of energy efficiency	- development of high yield technologies - initiatives which will reduce energy demand - development of technologies for utilizing residues
2. Enhancement of diversification/alternative energy sources	- initiatives which provide or strive for new energy sources - provide for flexibility in or adaptability of existing and new facilities
3. $ Saved internal to organization	- cost reduction devices - new technology to replace obsolete equipment
4. Impact on environment	- reduction of emissions into water and atmosphere - reduction of risk of nuclear accidents
5. Enhancement to *internal* technical capability & research profile	- provides training & develops expertise - provides technical resources (software, equipment, etc.) - builds linkages to external research community.
6. Enhancement to research Profile as viewed by the *external* community	- impact on research status among other utility companies - impact on profile abroad
7. Economic impact on external community	- job creation outside organization - $ savings to public & industry created by energy efficiency devices
8. Impact on nuclear performance	- influence on nuclear station maintenance, etc.

Chapter 10

RESOURCE CONSTRAINED DEA

10.1. INTRODUCTION

In certain performance measurement situations it is required to select a subset of alternatives from a given set of choices, and in a resource-constrained environment. When both multiple outputs and inputs are involved, the DEA model structure offers the opportunity to make choices based upon optimizing aggregate output relative to aggregate input.

In this chapter we examine two examples of such resource-constrained settings. The first involves choosing from a set of projects, a subset which is to be implemented. Each project is expected to make use of input resources of various types, to produce a set of outputs. In essence, our approach treats each subset of the projects that could feasibly be selected within the resource constraints as a single, composite project. These composite projects are then evaluated, by data envelopment analysis, against a 'production technology' defined by the available projects. In fact, evaluation and selection are combined in a single model by placing the data envelopment analysis model within a mixed-binary linear programming framework. This model is illustrated using Oral, Kettani and Lang's (1991) data on 37 R&D projects.

The second application involves selecting a set of best or preferred sites for planned retail facilities. Again, the model is concerned with choices in a resource-constrained setting.

10.2. PROJECT PRIORITIZATION: A RESOURCE-CONSTRAINED DEA APPROACH

10.2.1 Introduction

The application discussed here falls firmly into the multi-criteria decision making arena. It involves the selection, from a larger set of proposals, of a subset of projects to be undertaken. Individual projects are expected to produce benefits under a number of headings and, in so doing, will make use of resources under a number of headings. It is desired to establish a subset of projects that can be justified to all concerned as making the best use of the available resources.

This *prioritization* problem, in various forms, has received substantial attention over the past several decades. See Martino (1995). Our approach to the problem has its origins in Bunch et al. (1989) but is specifically related to that of Oral, Kettani and Lang (OKL) (1991). OKL's point of departure is identical to ours: the CCR DEA model.

In the interest of fairness to each of the proposed projects, OKL erect a rather complex multi-stage collective evaluation and selection model on this foundation. Our approach, which combines evaluation and selection in a single stage, remains substantially faithful to CCR DEA, and is somewhat less complex.

10.2.2 Modelling Preliminaries

A set $P(=\{1...k...|P|\})$ of project proposals is to be evaluated over a set $O(=\{1...j...|O|\})$ of outputs and a set $I(=\{1...i...|I|\})$ of inputs. Project k is characterized by the magnitudes of its outputs $y_{kj}(\geq 0)$ to be produced and its inputs $x_{ki}(\geq 0)$ to be consumed. There is a limit L_i on the quantity of input i available to the set of projects as a whole, and we assume that at least one project satisfies these limits. It is desired to select a subset of projects, $s^* \subset P$, which can be justified as making the best use of the available resources.

It is assumed that all the projects are, in principle, supportable; all would be undertaken in the absence of the resource constraints. It is also assumed that the projects are neither synergistic nor interfering, in the sense that, if both projects α and β were selected, the outputs thus produced would be the sum of their respective outputs and similarly for the inputs used.

If some function, $\theta_k = \theta(x_{k\ell},...,x_{ki},...,y_{k\ell},...,y_{kj},...)$ were available, such that it were possible to arrive at an 'objective' evaluation θ_k of each project, a net benefit say, this could be used to rank the projects. Further, s^*

could be obtained in a relatively straightforward manner. A natural representation for this situation might be as a binary knapsack problem along the following lines:

$$\max \sum_{k\in P} c_k\theta_k$$
$$\text{subject to}$$
$$\sum_{k\in P} c_k x_{ki} \le L_i \qquad i \in I \tag{10.1}$$
$$c_k \in \{0,1\} \qquad k \in P$$

Difficulties arise because of the non-availability of θ, or there might at least be disagreement among the various interested parties concerning its form and detail.

A DEA-based approach circumvents these difficulties by allowing each project proposal to evaluate itself, *relative* to all the projects under consideration. Essentially, each project k is allowed to rate itself as highly as possible via a kind of benefit/cost ratio:

$$h_k = (\sum_{j\in O} u_{kj} y_{kj}) / (\sum_{i\in I} v_{ki} x_{ki}),$$

by choosing the weights u_{kj} and v_{ki} to be applied to its outputs and inputs, respectively. The only restriction imposed is that no project p is allowed to receive a rating greater than 1 with those weights. This self-evaluation is achieved by solving the following linear program on behalf of each project k, as per Charnes et al. (1978). (Also see Charnes et al. (1991) for a formal treatment of situations where one or more of x_{ki}, y_{kj} might be zero.):

$$\max \; h_k = \sum_{j\in O} u_{kj} y_{kj} \qquad k \in P$$
$$\text{subject to}$$
$$\sum_{i\in I} v_{ki} x_{ki} = 1 \tag{10.2}$$
$$\sum_{j\in O} u_{kj} y_{pj} - \sum_{i\in I} v_{ki} x_{pi} \le 0 \qquad p \in P$$
$$u_{kj}, v_{ki} \ge 0, \; j \in O, \; i \in I$$

While self-evaluation in this way is entirely notional, there is an implicit fairness in the process. The ratings achieved depend only on the data for each project relative to the data for the other projects.

The values of h_k might now be used to rank the projects, but the problem of how to select a subset to support within the resource constraints persists. It is tempting to simply replace θ_k by h_k in the binary knapsack problem of (10.1). However, this would, in general, be misguided, as the following example indicates. Imagine the situation where there are a number of projects, each with a single output and a single input. Three of the projects, A, B and C, have the following values for $x_{.1}, y_{.1}$ and $h_.$, respectively:

A	400	400	1.00
B	300	225	0.75
C	100	26	0.26

It can easily be seen that from the viewpoint of the knapsack model, project A is inferior to a combination of projects B and C. However, the latter combination is obviously inferior to A, in terms of the quantity of output produced per unit of input, and should never be chosen in preference to A.

Clearly, while h_k *may* provide a meaningful ranking of the projects, to the limits of the discrimination available given an upper limit of 1, it is not appropriate to treat these values additively as in (10.1). In order to retain the apparent flexibility and fairness offered by a DEA-based approach, we combine evaluation and selection into a single prioritization model, as described in the following section.

Before proceeding, however, it is worth pointing out that our proposed model attempts to draw a compromise between what might be regarded as two opposing views of optimal selection in this context. One view is the traditional benefit/cost ratio approach to evaluating a set of choices (e.g. projects). This approach concentrates on the output per unit of input; project α is thus preferred to project β if the benefit/cost ratio of the former exceeds the latter. No direct consideration of budget limitations on the inputs is given at the evaluation stage; these must, somehow, be considered at the selection stage.

An alternative view is that typified by the usual mathematical programming approach where benefit/cost ratios are not a direct consideration; rather, satisfying budget constraints on the input resources while maximizing some measure of total output (benefit) is the goal. Said another way, if two groups of projects both meet the resource constraints (perhaps more than one such constraint), and yield equal aggregate benefits, we would be indifferent as far as the desirability of these two groups was concerned. From the benefit/cost viewpoint, however, the group with the smaller cost would be preferred. So, on the one hand (the benefit/cost approach), resources have a value and the less used, the better. On the other hand (the mathematical programming approach), we essentially assume resources have no value, except when we try to exceed their budgetary limits. What we propose herein is an approach to evaluation and selection that tries to capture both of these aspects.

10.2.3 A Prioritization Model

Given that the individual projects are independent, neither synergistic nor interfering, any subset s of the set P of projects can be thought of as a single, composite project. The outputs of this composite project Y_{sj} are the combined-by-addition outputs of its constituent projects; similarly for its inputs X_{si}. The focus of interest now, of course, is an evaluation of each composite project relative to the set of all such composite projects. This latter set we will call $\Pi(P)$ and is essentially the so-called power set of P (Halmos (1960)) (excluding the empty set ϕ). The individual projects constituting the highest rated composite(s), satisfying the resource constraints, are then candidates for selection. Thus, noting that $Y_{sj} = \sum_{k \in s} y_{kj}$ and $X_{si} = \sum_{k \in s} x_{ki}$, equation (10.2) becomes:

$$\begin{aligned}
&\max\ h_s = \sum_{j \in O} u_{sj} y_{sj} \qquad s \in \Pi(P)\\
&\text{subject to}\\
&\sum_{i \in I} v_{si} x_{si} = 1 \\
&\sum_{j \in O} u_{sj} y_{pj} - \sum_{i \in I} v_{si} x_{pi} \le 0 \qquad p \in \Pi(P)\\
&u_{sj}, v_{si} \ge 0 \qquad j \in O, i \in I
\end{aligned} \tag{10.3}$$

As the number of elements (composite projects) in $\Pi(P)$ is $2^{|P|} - 1$, which is large even for relatively modest $|P|$, (10.3) does not represent a practical proposition. However, as a first step toward practicality, the number of "$\le$" constraints in (10.3) can be reduced from $2^{|P|} - 1$ to $|P|$.

Imagine dividing the "$\le$" constraints in (10.3) into two groups: the first group is associated with the singleton subsets of P i.e. $\{1\}, \{2\} \ldots \{|P|\}$, while the second group is associated with the non-singleton subsets e.g. $\{1,2\}$, $\{1,3\}$, etc. (10.3) can then be written as:

$$\begin{aligned}
&\max\ h_s = \sum_{j \in O} u_{sj} y_{sj} \qquad s \in \Pi(P)\\
&\text{subject to}\\
&\sum_{i \in I} v_{si} x_{si} = 1 \\
&\sum_{j \in O} u_{sj} y_{pj} - \sum_{i \in I} v_{si} x_{pi} \le 0 \qquad p \in P\\
&\sum_{j \in O} u_{sj} Y_{qj} - \sum_{i \in I} v_{si} X_{qi} \le 0 \qquad q \in \Pi(P) - P'\\
&u_{sj}, v_{sj} \ge 0 \qquad j \in O, i \in I
\end{aligned} \tag{10.4}$$

where $P' = \{\{1\}, \{2\} \ldots \{|P|\}\}$, the set of all singleton subsets of P.

It is evident that any constraint in the second group of "$\le$" constraints is an additive combination of two or more constraints in the first group. Thus, if the constraints in the first group are satisfied, then, so must any constraint in the second group. Therefore, the second group contains only redundant constraints and can be removed. The basic prioritization model for composite projects thus becomes:

$$\max \; h_s = \sum_{j \in O} u_{sj} y_{sj} \qquad s \in \Pi(P)$$

subject to

$$\sum_{i \in I} v_{si} X_{si} = 1 \tag{10.5}$$

$$\sum_{j \in O} u_{sj} y_{pj} - \sum_{i \in I} v_{si} x_{pi} \le 0 \qquad p \in P$$

$$u_{sj}, v_{si} \ge 0 \qquad j \in O,\ i \in I$$

We now restrict the scope of the index s in the objective function of (10.5) by recognizing that, in this context, interest would be restricted to a particular subset of $\Pi(P)$. This subset S can be characterized by the following two conditions:

Condition (a): For all $s \in S$, the constraints on all resources are satisfied:

$$\forall i \in I,\ \sum_{k \in s} x_{ki} \le L_i,$$

Condition (b): For all $s \in S$, no project can be added without violating Condition (a):

$$\forall p \in P - s, \exists i \in I \text{ such that } \sum_{k \in s \cup \{p\}} x_{ki} > L_i.$$

Condition (a) is an obvious requirement while Condition (b) follows from the observation that all projects are supportable. Any composite project to which a further project could be added without violating Condition a) would be so augmented. Thus, the proposal is to look only at those composites that absorb at least one of the resources up to its usable limit; i.e. any amount of that resource left over is not sufficient to permit inclusion of another project. Then, within that subset of composites, one finds the composite whose aggregate benefit to aggregate cost ratio is maximized.

Rather than generating the set S explicitly, and subsequently evaluating each of its members via (10.5), we do so implicitly by placing (10.5) within a mixed-binary non-linear programming framework (10.6), below. Here, c_k is 1 if project k is included in the composite s^*, and 0 otherwise. Optimization now takes place over c_k, u_j and v_i.

$$\max_{(c_k, u_j, v_i)} \sum_{j\in O} u_j (\sum_{k\in P} c_k y_{kj})$$

subject to

$$\begin{aligned}
&\sum_{i\in I} v_i (\sum_{k\in P} c_k x_{ki}) = 1 \\
&\sum_{j\in O} u_j y_{pj} - \sum_{i\in I} v_i x_{pi} \le 0 && p \in P \\
&\sum_{k\in P} c_k x_{ki} + l_i = L_i \quad i \in I \\
&(1-c_k)x_{ki} + \mathrm{M}c_k + \mathrm{M}d_{ki} \ge l_i + 1/\mathrm{M} && k\in P, i \in I \\
&\sum_{i\in I} d_{ki} \le |I| - 1 && k \in P \\
&c_k, d_{ki} \in \{0,1\} && k\in P, i\in I \\
&u_j, v_i, l_i \ge 0 && j\in O, i\in I \\
&\mathrm{M}>>0
\end{aligned} \tag{10.6}$$

Before going on to linearize (10.6), some explanation is in order. The model now seeks the best evaluated subset satisfying Conditions (a) and (b), above. Obviously, Condition (a) translates directly into the constraints:

$$\sum_{k\in P} c_k x_{ki} + l_i = L_i,$$

where l_i is the slack in resource i. Condition (b) is a little more difficult but is implemented by the constraints:

$$(1-c_k)x_{ki} + \mathrm{M}c_k + \mathrm{M}d_{ki} \ge l_i + 1/\mathrm{M} \qquad k\in P,\ i\in I$$

$$\sum_{i\in I} d_{ki} \le |I| - 1 \qquad k\in P.$$

The effect here is to require that at least one of the resource slacks, l_i, be too small to allow another project into s^*. For a given $s \in S$, consider the first of these constraints for some $k \in s$, i.e. $c_k = 1$. The constraint is obviously satisfied because of the positive multiple of M on the left hand side. Now, consider the situation for some $k \notin s$, i.e. $c_k = 0$, and $x_{ki} \le l_i$. The corresponding first constraint can be satisfied by setting $d_{ki} = 1$, thus achieving a positive multiple of M on the left hand side. However, the effect of the second constraint is to ensure that at least one of the variables $d_{k1}, d_{k2} \ldots d_{k|I|}$ remains at zero. Hence, $\exists i \in I$ such that $x_{ki} \ge l_i + 1/\mathrm{M}$ as required.

While software capable of solving (10.6) is available, it can be linearized to bring it within the capability of more readily available mixed-binary linear programming software. This linearization involves the following changes of variables: $a_{kj} = c_k u_j$ and $b_{ki} = c_k v_i$.

Model (10.6) now becomes:

$$\max_{(c_k, a_{kj}, b_{ki}, u_j, v_i)} \sum_{(k\in P, j\in O)} a_{kj} y_{kj}$$

subject to

$$\begin{aligned}
&\sum_{(k\in P, i\in I)} b_{ki} x_{ki} = 1 \\
&\sum_{j\in O} u_j y_{pj} - \sum_{i\in I} v_i x_{pi} \le 0 \qquad p \in P \\
&\sum_{k\in P} c_k x_{ki} + l_i \le L_i \qquad i \in I \\
&(1-c_k)x_{ki} + \mathrm{M}c_k + \mathrm{M}d_{ki} \ge l_i + 1/\mathrm{M} \qquad k\in P,\ i \in I \qquad (10.7) \\
&\sum_{i\in I} d_{ki} \le |I| - 1 \qquad k \in P \\
&c_k, d_{ki} \in \{0,1\} \qquad k\in P,\ i\in I \\
&\left.\begin{array}{l} a_{kj} \ge 0 \\ a_{kj} \le \mathrm{M}c_k \\ u_j \ge a_{kj} \end{array}\right| k\in P,\ j \in O \\
&\left.\begin{array}{l} b_{ki} \ge 0 \\ b_{ki} \le \mathrm{M}c_k \\ v_i \ge b_{ki} \\ v_i \le b_{ki} + \mathrm{M}(1-c_k) \end{array}\right| k\in P,\ i\in I \\
&\mathrm{M} >> 0
\end{aligned}$$

where the two sets of constraints highlighted by the vertical bars serve to connect the new variables a_{kj}, b_{ki} to the original variables c_k, u_j and v_i.

Before applying this model in the next section, it is important to note that the fairness in evaluation implicit in (10.2) is retained in our prioritization model. Each project proposal thus has an equal right to participate in the definition of the 'production technology' as well as to combine with other projects to be evaluated against said technology and to be selected. This process depends only on the data for the projects relative to each other and on the available resources.

10.2.4 An Application

OKL (1991) demonstrate their approach to collective evaluation and selection in an application relating to the Turkish iron and steel industry. Here, 37 projects are available, each of which is predicted to provide benefits under five headings:

- *direct economic contribution* to the iron and steel sector through improved quality and productivity, cost reductions, etc.

- *indirect economic contribution* to sectors depending on the iron and steel sector through better quality, lower prices, etc.
- *technological contribution* through better use of imported technology, etc.
- *scientific contribution* in the sense of better use of existing scientific knowledge, advancing the body of scientific knowledge, etc.
- *social contribution* in terms of job creation, better working conditions, higher living standards, etc.

In achieving benefits under these headings, each project would require a budget allocation from a single monetary resource. The data for the projects are shown in Table 10-1.

The total resource available to the selected projects is 1000.00 units. The average resource requirement over the 37 projects is 67.99. It could therefore be expected that s^* would contain in the order of 15 projects.

On solving (10.7), using AMPL/CPLEX, with the data summarized in Table 10-1, $s^* = \{1, 6, 14, 15, 16, 17, 18, 23, 26, 27, 31, 32, 34, 35, 36, 37\}$ with a collective rating (h_s) of 0.700 and a total resource use of 962.8 units. For purposes of comparison, OKL's selected subset is identical except our projects 6 and 32 are replaced by 21 and 29, with a resource use of 964.7. We thus agree on 14 of the 16 projects selected. Their selected subset has a collective rating of 0.690 when evaluated by (10.5).

It should be emphasized that the weights u_j, v_i do not reflect any *a priori* judgments concerning their absolute or relative values. If it is considered important to reflect such judgments within the prioritization process, further constraints can be added in the manner of Thompson et al.'s (1986) 'assurance region' extension to the CCR DEA model. In general, we can consider (10.7) as augmented with a possibly empty set of constraints $\text{AR}(u_j, v_i)s$. These represent any restrictions on the weights and their inter-relationships that the decision maker(s) deem appropriate.

By way of sensitivity analysis for our solution to OKL's problem, we have experimented with various forms of assurance region augmentation to (10.7). As an illustration, $\text{AR}(u_j, v_i)s$ of the form:

$$u_{j1} \geq u_{j2} \geq u_{j3} \geq u_{j4} \geq u_{j5},$$

where $j^1, \ldots, j^5$, a permutation of the integers $\{1, \ldots, 5\}$, reflect weak orderings of the weights u_j. Taking all such weak orderings into account identifies a robust 'core' of 13 projects {1, 14, 16, 17, 18, 23, 26, 27, 31, 34, 35, 36, 37} which is invariably selected. It also identifies a 'margin' of 6 projects {6, 11, 15, 21, 29, 32} which are selected in various groups of 2 or 3 according to the specific ordering imposed on the weights.

Table 10-1. Data on 37 research and development projects relating to their expected performance on five criteria and their costs

R&D	Indirect	Direct	Technical	Social	Scientific	Proj. Cost
1	67.53	70.82	62.64	44.91	46.28	84.20
2	58.94	62.86	57.47	42.84	45.64	90.00
3	22.27	19.68	6.73	10.99	5.92	50.20
4	47.32	47.05	21.75	20.82	19.64	67.50
5	48.96	48.48	34.90	32.73	26.21	75.40
6	58.88	77.16	35.42	29.11	26.08	90.00
7	50.10	58.20	36.12	32.46	18.90	87.40
8	47.46	49.54	46.89	24.54	36.35	88.80
9	55.26	61.09	38.93	47.71	29.47	95.90
10	52.40	55.09	53.45	19.52	46.57	77.50
11	55.13	55.54	55.13	23.36	46.31	76.50
12	32.09	34.04	33.57	10.60	29.36	47.50
13	27.49	39.00	34.51	21.25	25.74	58.50
14	77.17	83.35	60.01	41.37	51.91	95.00
15	72.00	68.32	25.84	36.64	25.84	83.80
16	39.74	34.54	38.01	15.79	33.06	35.40
17	38.50	28.65	51.18	59.59	48.82	32.10
18	41.23	47.18	40.01	10.18	38.86	46.70
19	53.02	51.34	42.48	17.42	46.30	78.60
20	19.91	18.98	25.49	8.66	27.04	54.10
21	50.96	53.56	55.47	30.23	54.72	74.40
22	53.36	46.47	49.72	36.53	50.44	82.10
23	61.60	66.59	64.54	39.10	51.12	75.60
24	52.56	55.11	57.58	39.69	56.49	92.30
25	31.22	29.84	33.08	13.27	36.75	68.50
26	56.64	58.05	60.03	31.16	46.71	69.30
27	50.40	53.58	53.06	26.68	48.85	57.10
28	30.76	32.45	36.63	25.45	34.79	80.00
29	48.97	54.97	51.52	23.02	45.75	72.00
30	59.68	63.78	54.80	15.94	44.04	82.90
31	48.28	55.58	53.30	7.61	36.74	44.60
32	39.78	51.69	35.10	5.30	29.57	54.50
33	24.93	29.72	28.72	8.38	23.45	52.70
34	22.32	33.12	18.94	4.03	9.58	28.00
35	48.83	53.41	40.82	10.45	33.72	36.00
36	61.45	70.22	58.26	19.53	49.33	64.10
37	57.78	72.10	43.83	16.14	31.32	66.40

10.3. CHOICE OF DEA MODEL

The choice of the CCR DEA model as the basis for our prioritization model implies that its underlying empirical 'reference technology' or 'production possibility set' is suitable to our purposes. See, for example, Grosskopf (1986) and Maindiratta (1990).

An (empirical) production possibility set is a declaration of the totality of potential production possibilities that might plausibly be observed. In our case, this is based on the evidence of the finite collection of production possibilities that are *to be* observed. In a situation where there is a set $D = \{1...d...|D|\}$ of decision making units (DMUs) where DMU d has produced a vector of outputs $Y_d = (Y_{d1}...Y_{dj}...Y_{d|O|})$ from a vector of inputs $X_d = (X_{d1}...X_{di}...X_{d|I|})$, then the CCR production possibility set, $T^{CCR}(D)$, can be represented as:

$$\{(X,Y) \in \Re^{|I|+O|} \mid \sum_{d\in D} \lambda_d X_d \le X, \sum_{d\in D} \lambda_d Y_d \ge Y, \lambda_d \ge 0, \forall d \in D\} \quad (10.8)$$

We then identify the set D of DMUs in (10.8) with our set of composite projects $\Pi(P)$, where the latter obviously contains the individual projects as the singleton subsets of P. It can then be immediately observed that $T^{CCR}(\Pi(P))$ contains non-negatively scaled versions of all projects, individual and composite. This is implicit in (10.3), above, which is the starting point for our prioritization model.

The argument behind the derivation of model (10.5) from (10.3) can now be seen. Imagine the set of input/output vectors corresponding to the composites in $\Pi(P)$ divided into two subsets: those associated with the $|P|$ singleton subsets $\{1\},\{2\},...\{|P|\}$ (i.e. the individual projects themselves), and the remainder associated with the non-singleton subsets (i.e. the composites).

We can therefore write the vector $(\sum_{p\in\Pi(P)} \lambda_p X_p, \sum_{p\in\Pi(P)} \lambda_p Y_p)$ in (10.8) as

$$(\sum_{p\in P} \lambda_p x_p + \sum_{q\in(\Pi(P)-P')} \lambda_q X_q, \sum_{p\in P} \lambda_p y_p + \sum_{q\in(\Pi(P)-P')} \lambda_q Y_q) \quad (10.9)$$

where $P' = \{\{1\},\{2\},...\{|P|\}\}$. Now, with the convention that an index q identifying a composite in (10.9) also identifies the subset of projects comprising that composite, any $\lambda_q \ne 0$ in (10.9) can be set to zero by the algorithm:

$$\forall p \in q \lambda_p \to \lambda_p + \lambda_q.$$

This follows from the way that input/output vectors corresponding to composites are constructed, i.e., by addition of the input/output vectors of the projects themselves. Repeated application of the above algorithm for all $\lambda_q \ne 0$ would serve to drive all λ_q to zero. $T^{CCR}(\Pi(P))$ is essentially equivalent to $T^{CCR}(P)$. The latter is thus capable of representing the former.

A key feature of the CCR reference technology, by virtue of the unbounded (from above) multipliers λ_p, is that constant returns to scale are assumed. We regard this as appropriate here for two main reasons. Firstly, there seems little merit in rewarding projects for being relatively efficient technically but at an inappropriate scale. See Banker et al. (1984) in terms of their proposed conversion of inputs into outputs. Secondly, the effect of the resource constraints will ensure that the composites in S will operate at a similar scale, subject only to the granularity inevitable in combining discrete projects.

However, if it were required to take scale aspects into account, this would typically be done via the Banker, Charnes and Cooper (BCC) (1984) production possibility set, $T^{BCC}(\Pi(P))$. This differs from $T^{CCR}(\Pi(P))$ in that, as well as the lower bounds on the multipliers λ_p, there is a 'convexity' constraint in the form of $\sum_{p\in\Pi(P)}\lambda_p = 1$ on their sum. It can easily be seen that and $T^{BCC}(\Pi(P))$ and $T^{BCC}(P)$ are not equivalent; the composite project equivalent to the sum of all the individual projects is in the former, by definition, but not in the latter.

Importantly, Kao (1998) shows that $T^{BCC}(\Pi(P))$ is equivalent to the Koopmans production possibility set (see Grosskopf (1986)) $T^{Koopmans}(P)$ which, itself, differs from $T^{CCR}(P)$ by the incorporation of upper bounds $\lambda_p \leq 1$ on the multipliers. It is a straightforward matter to modify (10.5) and, hence, (10.6) and (10.7), to implement $T^{Koopmans}(P)$ rather than, $T^{CCR}(P)$, if desired.

10.4. SELECTING SITES FOR NEW FACILITIES

10.4.1 Introduction

Site selection for facilities is an important problem that has been studied extensively and reported on widely in the literature. Among the numerous applications are location of facilities such as fire stations, ambulance depots, and police stations. See, for example Savas (1969). These applications are aimed primarily at minimizing the distances between supply and demand centers. A related class of site selection problems pertains to coverage; an example is the deployment of health care clinics in rural areas (Eaton et al. (1982), and Calvo et al. (1973)). Here, the objective is to choose sites for clinics so as to optimize some function of total output (e.g. patients served), and generally in a resource constrained setting. In this situation, travel distance (or time) by the consumer is not directly an objective, but may be imposed in the form of a constraint.

Many existing site selection models tend to view situations from a uni-dimensional perspective. In Eaton et al. (1982) , for example, total patients served is the output of interest as opposed to defining different classes of outputs—outpatient clients, critical care patients, expectant mothers, etc. In this multi-output situation, one type of patient may impose different resource requirements than another type. As a result, this multi-dimensional setting is complicated by the fact that often there is no well defined rate of substitution between the different output types. One generally cannot say, for instance, that the resource requirements for one outpatient is equal to one-third the requirements for a critical care in-patient. Therefore, the multi-output setting cannot be converted to a single output situation. On the resource or input side, a similar situation can prevail. Different classes of inputs (e.g., staff types, equipment, physical facilities) may exist, for which there is not a common rate of exchange. Specialized equipment (e.g. for radiation treatments) will be used for only certain patient types.

As a departure from the uni-dimensional philosophy of traditional set covering approaches to site selection, a number of authors, including Fisher and Rushton (1979), have advocated the development of analytical techniques for handling the multi-criteria nature of the problem. One such multi-criteria methodology is the data envelopment analysis (DEA) technique. Thompson et al (1986) presented one of the earliest DEA applications involving the evaluation of sites for a high-energy physics lab in Texas. Later, Desai and Storbeck (1990) looked at the concept of relative spacial efficiency when two measures of access are involved, namely total travel distance and least number of people not covered. Balakrishnan et al. (1994) apply a two stage technique to choosing sites for retail outlets. In the *first* stage, a set of location—covering scenarios or configurations is generated, with each configuration satisfying given adequacy requirements. Then, in the *second* stage, viewing a configuration from a multi-criteria standpoint, DEA is applied to generate a score for that configuration. Athanassopoulos and Storbeck (1995) provide a comparative evaluation of DEA and the free disposal hull (FDH) method in the context of site selection. Again, they generate a candidate set of configurations of units, and then treat those configurations as the decision making units.

In the current section we demonstrate how DEA can be used to select sites for facilities in a *resource constrained setting*. As a practical setting for the model development, we examine the problem of selecting sites for new branches of a retail housewares and hardware store chain. We present a variation of the DEA model for choosing a best subset of a set of potential sites. This variation builds on the model of Section 10.2. This model is used to do a detailed case analysis of the selection of sites for retail stores.

10.4.2 Site Selection for a Retail Store Chain

A certain national chain of stores sells a wide range of houseware and hardware products. Products can generally be classified into a few major groups: (a) self-service housewares, (b) self-service hardware, (c) counter sales of automotive parts, (d) counter sales of electronic devices (televisions, stereos, VCRs, etc.), (e) furniture and appliance sales. The chain also operates a catalogue shopping arm. The company currently operates a set of 15 stores across the country, and wishes to select a set of new locations to expand its operation.

In selecting a location for a store, it can be argued that the merits of this location would be judged relative to the outputs generated (e.g., sales of the different classes of products) versus the inputs consumed (resources needed to support the facility). In simplistic terms, it is the profitability of the location that would determine its relative desirability. The difficulty which can arise in a setting such as this is that one type of product can place different demands on resources than another type. For example, self service products require little staff involvement, versus sales of furniture and appliances or even automotive parts where product knowledge is a prerequisite on the part of the sales person. Thus, it is necessary to have a forecast of sales potential in *different product classes* as opposed to simply a forecast of *aggregate* sales.

While it is the case that for each potential location, a forecast of aggregate sales dollars is needed, some caution must be taken in obtaining this forecast. Arguably, one could develop a forecast for each stock keeping unit (SKU) within a class, either in the form of monetary sales or units of the product. This could be done in the form of a regression model with independent variables such as population size, customer demographics, etc. For the stores in question, however, each class of products contained a large number of SKUs. Moreover, sales figures for a sample of SKUs examined showed a high degree of variability across the stores, and poor fits result from any of the models attempted.

Forecasting in this situation, for a large class of items, resembles the problem of developing aggregate production forecasts for workforce planning purposes in manufacturing. Essentially, one converts a collection of products into a single pseudo-product, representing the production hours consumed in each period. One then projects into the future, obtaining a forecast of upcoming labor needs. The analogy to aggregate planning in the current setting, would be to view net monetary sales (sales revenue less product cost), as the output for a product class. It would then appear that it is net dollar sales that one should attempt to forecast for any new proposed location. Here, however, net sales is highly variable among the existing locations, and appears to have little correlation to any of the logical predictor variables (population size, customer demographics, etc.) This may be due, in large part, to the fact that the selling price charged for an individual SKU can vary widely from one store location to another, due to promotions offered at the store level, discounts obtained when the product was purchased, etc. As well, the proportional mix of SKU volumes within a class can vary between stores. An examination of the average price per unit of product in a class (across the analysis sample of stores) showed a high degree of variability, thus contributing to the high variance in total sales revenue.

For aggregate sales planning purposes, the most predictable variable was the *total units of products sold.* As will be seen in a later section, this variable is highly correlated to certain predictor variables that are available for the proposed locations. Hence, it is this variable which we forecast. It is notable that if operational efficiency were at issue here, it is precisely this total units variable that one would choose to represent outputs, since it reflects staff workload. Workload is more a function of the number of transactions than of the profitability of those transactions. In our case, however, it is profitability, or return on investment, that is more correctly the issue in evaluating potential locations. Clearly one can represent aggregate net dollar sales as the product of total units sold and average profit per unit. Specifically, if y_{rk} is the forecast of total demand (in units) for product group r at site k, and if μ_r is the average net sales (selling priceless cost) per unit of product in that group, then the expected aggregate net sales from site k is $\sum_{r\in R} \mu_r y_{rk}$. Here R denotes the set of product groups. However, as discussed above, since the average price, (hence, average net sales), is highly variable from one location to another, μ_r is not explicitly available, although bounds are obtainable. It is this feature of the revenue function which points to DEA as a model structure for selecting sites. We discuss this below. On the input side there are two types of factors to be considered in terms of servicing customer demand (the outputs).The first type of input or influence is the resources available. This could be viewed in various ways, but the two most logical forms of resource are:

1. Initial capital outlay (construction and/or redecorating expenses), which is influenced by the size of the facility;
2. Annual operating expenses, including salaries, utilities, rental, etc.

Obviously, one could separate salaries in the form of total FTE staff numbers, but staff mix becomes an issue and is a function of the product mix problems discussed earlier.

A second level of influence factors are those that pertain to customer demographics, as well as those involving the influence of competition. To an extent these factors are part of the forecasting model for aggregate unit sales, although, for reasons to be discussed in a later section, the regression model may not include all such variables. For this reason, certain (nondiscretionary) factors may be included in the DEA model, rather than directly within the forecasting model.

Again, in notational terms if x_{ik} denotes the amount of input or factor i available at store location k, and v_i denotes the price or value associated with one unit of input x_{ik}, then the total resources dedicated to location k is $\sum_{i\in I} v_i x_{ik}$. Here I denotes the set of all inputs. This expression is somewhat less transparent than is true of weighted outputs $\sum_{r\in R} \mu_r y_{rk}$, in that we are mixing economic and non-economic factors on the input side. (In the application to be discussed, a single nondiscretionary variable, *level of competition*, is added on the input side.) Thus, it is useful to view the v_i merely as multipliers, and not directly as prices in the same sense that the μ_r can be interpreted.

10.4.3 A DEA Based Model

Consider the situation where K_1 existing facilities are in place, and let $\{y_{rk}\}_{r\in R}$ denote the R-component vector of outputs produced by facility k. Let $\{x_{ik}\}_{i\in I}$ denote the I-component vector of inputs. In the site selection situation, y_{rk} would denote the number of units sold of product group r at store k. Similarity, x_{ik} represents resource type i consumed by k, or is a nondiscretionary variable, depending on the value of i.

Suppose that K_2 potential sites are being considered as locations for new stores, and assume that a *forecast* is available for the numbers of units $\{y_{rk}\}_{r\in R}$ of product in groups $r \in R$. In the section to follow we discuss the development of such forecasts. On the input side, the competition variable x_{1k} for any location k is non-discretionary (i.e., it is a given value). The other two inputs x_{2k} = operating budget, and x_{3k} = capital outlay when establishing the store at the site, are at the discretion of the company. Thus, these latter two are *decision variables*. If a store is to be established at location k, the values assigned to x_{2k} and x_{3k}, will clearly influence the performance ratio of the store. Normally, the performance score for a site k would be expressed as $\sum_{r\in R} \mu_r y_{rk} / \sum_{i\in I} v_i x_{ik}$, computed relative to all other existing facilities, as well as newly created ones. In our case, however, with both discretionary inputs (DI) and nondiscretionary inputs (NDI), a proper form of the ratio is $[\sum_{r\in R} \mu_r y_{rk} - \sum_{i\in \text{NDI}} v_i x_{ik}] / \sum_{i\in \text{DI}} v_i x_{ik}$.

In a resource constrained setting the problem of interest is how to choose that subset S of the potential new sites $\{K_1+1, K_1+2, \ldots, K_1+K_2\}$ which will yield the greatest aggregate benefit to the organization, while not violating these resource limits. This idea is similar to the problem of project prioritization discussed in the previous sections as per Cook and Green (2000). The important difference here is that, unlike the project selection problem where the x_{ik} were *fixed* resource requirements (if a project is to be implemented), the x_{ik} in the current setting are decision variables that can take on values to be determined by the organization.

We let $I_1 \subseteq I$ denote that subset of (discretionary) inputs for which bounds are to be imposed. In the site selection problem, I_1 would generally consist of the two resource inputs – operating and capital expenditures. Correspondingly, $\overline{I}_1$ denotes the compliment of I_1. As well, let K_1 denote the existing set of DMUs $\{1,\ldots,K_1\}$, and K_2, the set of potential new sites $\{K_1+1,\ldots,K_2\}$. For any subset S, we represent the performance measure of S as the solution to the optimization problem:

$$\max \sum_{k\in S}[\sum_{r\in R}\mu_r y_{rk} - \sum_{i\in \overline{I}_i} v_i x_{ik}] / \sum_{k\in S}\sum_{i\in I_i} v_i x_{ik} \qquad (10.10a)$$

subject to

$$[\sum_{r\in R}\mu_r y_{rk} - \sum_{i\in \overline{I}_i} v_i x_{ik}] / \sum_{i\in I_i} v_i x_{ik} \le 1, \qquad k \in K_1 \quad (10.10b)$$

$$\overline{\theta}_1 \le [\sum_{r\in R}\mu_r y_{rk} - \sum_{i\in \overline{I}_i} v_i x_{ik}] / \sum_{i\in I_i} v_i x_{ik} \ge \overline{\theta}_2, \qquad k \in S \quad (10.10c)$$

$$\sum_{k\in S} x_{ik} \le C_i, \qquad i \in I_1 \quad (10.10d)$$

$$x_{ik} \ge L_i, i \in I_1, \qquad k \in S \quad (10.10e)$$

$$x_{ik} \ge 0, i \in I_i; \mu_r, v_i \ge \in, \qquad \text{all r, i} \quad (10.10f)$$

As discussed in Cook and Green (2000), when *composites* of DMUs (S is a composite) are being considered, the appropriate production technology is that provided by the CRS model of Charnes et al. (1985). Essentially one needs to assume that for any DMU, multiples of that DMU are also in the production possibility set. Recall, that the primal problem for CRS allows any $\lambda_j \ge 0$, whereas for the VRS model of Banker et al. (1984), one must operate within the convex hull of existing DMUs $(\sum \lambda_j = 1)$, meaning that composites may be beyond the bounds of the production possibility set. Hence, for composite considerations, CRS is the appropriate technology. Moreover, since the stores considered in the application herein, are of a relatively comparable size, non-constant returns to scale were deemed to be a non-issue.

Here, the lower bound $\overline{\theta}_1$ in (10.10c) is to be selected by management. It may, for example, be decided that the minimum performance level for any new site $k \in S$ must be at least $\overline{\theta}_1 = 80\%$. The upper limit of $\overline{\theta}_2$ for every store $k \in S$ may be taken as any value less than unity. We do not choose $\overline{\theta}_2 = 1$, as this would permit a yet to be established facility to have an arbitrarily small level of each resource ($> L_i$), possibly rendering all existing DMUs inefficient. As well, since output figures (total units of products by class) are estimates, their actuals may, in fact, exceed these levels, potentially putting proposed sites beyond the frontier of known performance. We therefore propose that only *known facilities* should define the production frontier.

Restrictions (10.10d) provide for limits C_i on the amount of resource available of input type i. Obviously, a limit on any nondiscretionary input may not have meaning. Constraints (10.10e) impose minimum requirements L_i, on resource commitments x_{ik} for established new sites (those where stores are to be placed). In the case of capital , L_i may be dictated by the minimum size facility that would be considered.

It should be pointed out at this stage that optimization of a ratio of aggregate output to aggregate input may result in only one site being selected, that is, the cardinality of S may be 1. In most instances this would mean that the constraints on resources C_i would play little or no role in the optimization process (except in the rare circumstance that some particular site required extensive resource input). The idea, however, is to select sites in a manner which makes the best use of available resources. This means that any set of sites S to which an additional site could be added, without violating the resource constraints, would be so augmented.

To ensure consideration is given only to subsets of sites to which no additional sites can be added, we first replace (10.10d) by the equivalent expression

$$\sum_{k \in S} v_i x_{ik} + v_i s_i = v_i C_i, i \in I_1 \tag{10.11}$$

where s_i is the slack variable corresponding to constraint $i \in I_1$ in (10.10d).

Next, to implicitly allow for consideration of all possible subsets S, we introduce binary decision variables d_k where

$$d_k = \begin{cases} 1 & \text{if a store is to be located at } k \\ 0 & \text{otherwise;} \end{cases}$$

We now replace y_{rk} by $d_k y_{rk}$, and the decision variable x_{ik} by $d_k x_{ik}$. Thus, any given subset of sites S corresponds to a particular set of positive d_k variables.

Finally, to ensure that no additional site can be added to a set S under consideration, it is necessary to require that at least one of the slack variables in (10.11) be too small to allow for enough of the corresponding resource to be committed to any new site. Specifically, any feasible set of sites S must be such that for at lease one input i, there is not sufficient resource s_i remaining (after the allocation to sites in S is made), to support any additional site from the complement of S. To accomplish this, introduce binary variables $r_{ik}, i \in I_1$, and $k \in K_2$, and add the constraints

$$v_i s_i \le v_i L_i - 1/M + M d_k + M r_{ik}, i \in I_1, k \in K_2 \tag{10.12}$$

$$\sum_{i \in I_1} r_{ik} \le |I_1| - 1, k \in K_2 \tag{10.13}$$

We can see from (10.12), that since any site $k \in S$ has $d_k = 1$, the constraints will be clearly satisfied for all i. For $k \in \overline{S}, d_k = 0$, meaning that either $v_i s_i \leq v_i L_i - 1/M$ or $r_{ik} = 1$. Constraint (10.13) permits only at most $|I_1| - 1$ of these binary variable r_{ik} to be unity, meaning that for any non-selected site k, *at least* one of the inputs i must have a (weighted) slack $v_i s_i$ strictly less than the (weighted) lower bound $v_i L_i$.

For any given $\{d_k\}$ and $\{x_{ik}\}$ (i.e, if these variables were fixed), it is noted that the resulting linear fractional programming problem (10.10)-(10.13) could be replaced by a linear programming equivalent. Since this non-fractional equivalent of (10.10)-(10.13) holds regardless of the values of d_k and x_{ik} (assuming they are not all zeros), then, the above problem can be converted to the form

$$\max \sum_{k \in K_2} [\sum_{r \in R} d_k \mu_r y_{rk} - \sum_{i \in \overline{I}_i} d_k v_i x_{ik}]$$

subject to

$$\sum_{k \in K_2} \sum_{i \in I_1} d_k v_i x_{ik} = 1$$

$$\sum_{r \in R} \mu_r y_{rk} - \sum_{i \in I} v_i x_{ik} \leq 0, \qquad k \in K_1$$

$$\sum_{r \in R} d_k \mu_r y_{rk} - \sum_{i \in \overline{I}_1} d_k v_i x_{ik} - \overline{\theta}_2 \sum_{i \in I_1} d_k v_i x_{ik} \leq 0, \qquad k \in K_2$$

$$\sum_{r \in R} d_k \mu_r y_{rk} - \sum_{i \in \overline{I}_1} d_k v_i x_{ik} - \overline{\theta}_1 \sum_{i I_1} d_k v_i x_{ik} \geq 0, \qquad k \in K_2$$

$$\sum_{k \in K_2} d_k v_i x_{ik} + v_i s_i = v_i C_i, \qquad i \in I_1$$

$$d_k x_{ik} \geq d_k L_i, \qquad i \in I, k \in K_2 \qquad (10.14)$$

$$v_i s_i \leq v_i L_i - 1/M + M d_k + M_{r_{ik}}, \qquad i \in I, k \in K_2$$

$$\sum_{i \in I_1} r_{ik} \leq |I_1| - 1, \qquad k \in K_2$$

$$x_{ik} \geq 0, \qquad i \in I_1, k \in K_2$$

$$\mu_r, v_i \geq \in, \qquad \text{all r, i}$$

$$d_k \in \{0,1\} \qquad \text{all k}$$

To facilitate writing this as a mixed binary linear problem, we first make the observation that if we add the constraints

$$x_{ik} \leq M d_k, k \in K_2, i \in I_1 \qquad (10.15)$$

then, for $i \in I_1$, the product $d_k x_{ik}$ can be replaced by x_{ik} wherever it appears. Next we make the changes of variables

$$a_{rk} = \mu_r d_k, \forall r, k \in K_2$$
$$b_{ik} = v_i d_k, \forall i, k \in K_2$$
$$c_{ik} = v_i x_{ik}, i \in I_1, k \in K_2$$
$$e_i = v_i s_i, i \in I_1, k \in K_2$$

To connect a_{rk} with μ_r and d_k, yet avoid defining the transformation explicitly as (nonlinear) constraints, we impose the restrictions

$$0 \le a_{rk} \le Md_k$$
$$a_{rk} \le \mu_r \le a_{rk} + M(1 - d_k).$$

From this, we see that $d_k = 0 \Rightarrow a_{rk} = 0$ and $\mu_r \ge 0$. If $d_k = 1$ then $\mu_r = a_{rk}$.

Similarly, the definition of b_{ik} implies

$$0 \le b_{ik} \le Md_k$$
$$b_{ik} \le v_i \le b_{ik} + M(1 - d_k).$$

Finally, in replacing $v_i x_{ik}$ by c_{ik}, for $i \in I_1$, we ultimately compute x_{ik} as c_{ik}/v_i, after the problem is solved. As well, slacks s_i are determined from e_i/v_i.

We note that condition (10.15) becomes (after multiplying through by v_i)

$$c_{ik} \le Mb_{ik}, i \in I_1 k \in K_2 \tag{10.16}$$

With the above change of variables and linking constraints, the required mixed integer formulation of our site selection problem is:

$$\max \sum_{k \in K_2} [\sum_{r \in R} a_{rk} y_{rk} - \sum_{i \in \bar{I}_1} b_{ik} x_{ik}]$$

subject to

$$\sum_{k \in K_2} \sum_{i \in I_1} c_{ik} = 1$$

$$\sum_{r \in R} \mu_r y_{rk} - \sum_{i \in I} v_i x_{ik} \le 0, \qquad k \in K_1$$

$$\sum_{r \in R} a_{rk} y_{rk} - \theta_2 \sum_{i \in I_1} c_{ik} - \sum_{i \in \bar{I}_1} b_{ik} x_{ik} \le 0, \qquad k \in K_2$$

$$\sum_{r \in R} a_{rk} y_{rk} - \theta 1 \sum_{i \in I_1} c_{ik} - \sum_{i \in \bar{I}_1} b_{ik} x_{ik} \ge 0, \qquad k \in K_2$$

$$\sum_{k \in K_2} c_{ik} + e_i = C_i v_i, \qquad i \in I_1$$

$$c_{ik} - L_i b_{ik} \ge 0, \qquad i \in I_1, k \in K_2$$

$$v_i L_i - e_i + Md_k + Mr_{ik} \ge 1/M, \qquad i \in I_1, k \in K_2$$

$$\sum_{i \in I_1} r_{ik} \le |I_i| - 1, \qquad k \in K_2 \qquad (10.17)$$

$$a_{rk} - Md_k \le 0, r \in R, \qquad k \in K_2$$

$$\mu_r - a_{rk} + Md_k \le M, r \in R, \qquad k \in K_2$$

$$\mu_r - a_{rk} \ge 0, r \in R, \qquad k \in K_2$$

$$b_{ik} - Md_k \le 0, \qquad i \in I, k \in K_2$$

$$v_i - b_{ik} + Md_k \le M, \qquad i \in I, k \in K_2$$

$$v_i - b_{ik} \ge 0, \qquad i \in I, k \in K_2$$

$$c_{ik} - Mb_{ik} \le 0, \qquad i \in I_1, k \in K_2$$

$$a_{rk}, b_{ik}, c_{ik}, e_i \ge 0, \mu_r \ge \in, v_i \ge \in, r \in R, \qquad i \in I, k \in K_2$$

$$d_k, r_{ik} \in \{0,1\}, \qquad i \in I, k \in K_2$$

10.4.4 An Application

The application to which we apply the model developed above pertains to the selection of prospective sites for a set of new home building stores, as described in Section 1. The data in Table 10-2 corresponds to the existing fifteen such stores. Associated with each store is the annual operating budget in units of $1,000. So, for example, the annual cost of operating store #1 is $750,000. To permit consideration of the capital cost associated with existing locations, the most practical approach seemed to be to combine its impact with operating cost. In this particular organization, the capital outlay for any given store was either borrowed funds or was viewed as such. Hence, interest and amortized principal (25 year amortization period) have been added to what would normally be considered as operating expenditures, namely salaries, wages, taxes, insurance, and all utilities. The figures displayed in the last column of Table 10-2 include the interest and annual principal consideration pertaining to the capital costs of stores currently operating. It must be noted that in some instances (e.g. stores 3, 7 and 9), the initial borrowed funds for construction have been repaid. In these cases the operating cost figures have been increased to reflect what the real (current) costs in annual interest and principal would be if those expenses were still in effect. For stores with ongoing capital costs, all figures have been adjusted to reflect current rates.

As discussed earlier, aggregate sales revenue among the 15 existing stores was found to be highly variable. This is due primarily to the wide variation in net profit per unit of any given product, but is, as well, a function of the fact that the proportional distribution of SKU sales volumes is somewhat variable from store to store. This latter aspect can likely be attributed to customer taste and level of affluence. The most predictable variable appears to be aggregate units sold, when the large numbers of SKUs were combined. We have chosen here to divide total sales volumes into two major classes, namely: Class 1 - furniture and appliances, where a high degree of floor sales effort is required; and Class 2 - all other sales (generally necessitating minimal floor staff involvement). These two sales figures appear as columns 1 and 2 in Table 10-2.

Columns 3, 4, and 5 in Table 10-2 are demographic factors. Competition records the number of direct competitors within a two-mile radius of the store in question. These competitors are retail establishments that sell a large proportion of the product lines carried by the store in question. Hence, store #1 has 3 direct competitors in the same vicinity. Column 4 displays the number of single family dwellings per thousand persons in the metropolitan area where a store is located. This variable was chosen as a surrogate for disposable income, the rationale being that homeowners may reflect a higher level of affluence than non-homeowners, and are, thus, more likely to purchase higher value items. Column 5 provides figures for total population (in thousands) in the store vicinity. Since most of the stores are located in smaller cities and towns, the vicinity is generally defined as the entire city. In the case of stores 1, 4, and 6, however, which are all located in the same large city, the "vicinity" in each case is a two-mile radius of the store.

In formulating a predictive model, the correlation matrix provides some insight into the connections among the variables of Table 10-2. It is noted that housing and population are highly correlated (.769), hence the product "Popsing" of these (total houses in the store vicinity) was computed; the revised correlation matrix is displayed in Table 10-3. Arguably, although competition is not highly correlated to sales, it would appear to be appropriate to include it in the analysis. The resulting regression model is given by:

$$\text{Sales } 1 = 41.5 + .05\ (5.93)\ \text{House} + .73\ (.99)\ \text{Comp}$$

The t-values are shown in brackets. It is noted that the competition variable is not only insignificant, but, as well, assumes the incorrect sign (it is negatively correlated to sales as per Table 10-3). This is very likely caused by the relatively low correlation between competition and Sales 1, versus the higher correlation between Popsing and Comp.

A plausibly more appropriate model for forecasting sales is to base the latter on total housing alone, namely

$$\text{Sales } 1 = 30.96 + 0.054\ (5.58)\ \text{Popsing}$$

The total housing variable has a t-value of 5.58, and the overall R^2 value is 79.5%. This model has been used to generate the Sales 1 figures of Table 10-3, for the six potential locations.

Correspondingly, the forecasting model for Sales 2 is give by

$$\text{Sales 2} = 47.46 + 0.038\ (3.45)\ \text{Popsing}$$

which has an R^2 -value of 66.2%. Again the predicted Sales 2 values for the six proposed locations are displayed in Table 10-4.

Table 10-2. Data on Existing Stores

Store	Outputs Furn/Appl Sales 1 (100 units)	Other Sales 2 (1000 units)	Inputs(Predictors) Competition	# Single homes (100)	Total Pop. Vicinity (1000)	Annual Operating Cost ($1,000)
	1	2	3	4	5	6
1	58	73	3	15	20	750
2	21	46	4	7	15	490
3	49	55	3	13	22	680
4	63	68	2	17	30	620
5	57	70	2	15	25	730
6	62	65	1	18	23	860
7	41	52	3	10	18	520
8	35	46	3	8	12	390
9	39	49	3	9	11	490
10	29	45	4	7	12	420
11	33	40	4	6	14	550
12	48	65	2	14	28	730
13	52	68	1	18	20	880
14	65	65	1	20	43	960
15	51	72	2	17	18	820

Table 10-3. Correlation Matrix

	Comp	Single	Pop	sales1	sales 2	Popsing
Comp	1					
Single	-0.45557	1				
Pop	0.014427	0.768513	1			
sales1	-0.35883	0.928864	0.764681	1		
sales1	-0.46301	0.907664	0.607424	0.859388	1	
Popsing	-0.507	0.870882	0.970757	0.891511	0.814325	1

Table 10-4. Data on Potential Sites

Site #	Predicted Sales 1 (100 units)	Predicted Sales 2 (1000 units)	Competition	# Single homes	Population
1	54	64	3	17	25
2	44	57	4	12	20
3	74	78	2	20	40
4	36	51	5	8	12
5	78	81	1	25	35
6	51	62	3	15	25

The DEA Analysis

For the DEA analysis, Sales 1 and Sales 2 were the outputs used, with annual operating expenditure (which includes annualized capital outlay), and competition being the inputs.

As discussed earlier, competition should be seen as a nondiscretionary input, and was treated as such in the analysis. Arguably, in a resource allocation setting such as this, one should only be concerned with the profitability of the various sites, hence nondiscretionary variables should not really be a consideration in the DEA analysis. However, it can be claimed that if the competition variable *had been* kept as part of the forecasting model as discussed above, it is likely that the estimated sales figures would have been different from those currently available. Specifically, it can be argued that those sites with a high competition may have overstated sales figures, and those with low competition, are possibly understated. Thus, if the DEA score is presented as in (10.10a), we indirectly make the nondiscretionary variable part of the forecast.

As indicated above, the capital cost component has been included as part of the annual operating budget, hence the DEA analysis provides for a single discretionary input. This avoids, at the same time, undesirable weight differences (the exchange rates) that might occur if separate operating and capital inputs were used.

The problem of the high variability in net sales revenues described earlier, is addressed here by way of an analysis of average prices across the 15 stores for the two classes of products. The ranges of average prices were:

$$\text{Sales 1} \rightarrow (\$355, \$521)$$
$$\text{Sales 2} \rightarrow (\$5.75, \$9.35)$$

Given that Sales 1 figures in Tables 10-2 and 10-4 are expressed in hundreds of units and, those for Sales 2 are in thousands of units, these two ranges lead to the assurance region restrictions

$$3.8 \le \frac{\mu_2}{\mu_1} \le 9,$$

or more appropriately the pair of constraints

$$\mu_2 - 9\mu_1 \le 0$$
$$\mu_2 - 3.8\mu_1 \ge 0.$$

Clearly, other restrictions can be imposed here. For example, although the model given in the previous section has no specific provision for the extent of coverage provided by the selected sites, various modelling restrictions can facilitate this. One example would be to impose the constraint $\sum_{k \in K_2} d_k \ge N$, where N is a lower limit on the number of sites to be selected. Such a condition is easily implemented in the optimization model, and can permit management to evaluate various scenarios.

In this particular application, the lower and upper limits on the efficiency of any composite of sites were chosen as θ_1 = 80% and θ_2 = 98%, respectively. Recall that with the constant returns to scale model, composites of proposed store sites legitimately fall within the production technology. Hence, these composites are evaluated against the existing CRS frontier as if they were actual operating units. This means that if sites 1 and 5, for example, were combined as a single entity, and considered as a DMU, the efficiency score would be 84.3%. While the objective was to obtain a single best selection of the available six sites, as described in Table 10-4, management was interested in looking at a range of options, since certain *coverage* issues were also at stake.

Table 10-5. Feasible Combinations of Sites, and Their Corresponding Efficiency Scores

	Proposed Sites						
	1	2	3	4	5	6	
	d_1	d_2	d_3	d_4	d_5	d_6	Efficiency Score (%)
1.	x				x		84.3
2.		x		x			81.8
3.			x		x		87.6
4.					x	x	82.4
5.	x	x	x				93.5
6.	x	x		x			84.1
7.	x	x			x		91.7
8.	x	x				x	88.7
9.	x		x			x	92.6
10.	x		x		x		96.8
11.	x		x	x			86.3
12.	x			x	x		85.8
13.	x			x		x	85.3
14.	x				x	x	93.5
15.		x			x	x	90.5
16.		x	x	x			86.3
17.		x	x		x		93.7
18.		x	x			x	88.9
19.		x		x		x	84.1
20.		x		x	x		86.8
21.			x	x	x		88.6
22.			x	x		x	86.5
23.			x		x	x	94.9
24.				x	x	x	87.0

Table 10-5 presents the outcomes for all possible combinations of the 6 sites where the available annual operating budget is $2,250,000. (Any combinations not shown were not feasible). It would appear that the best combination of sites is combination #10 which composes sites 1,3 and 6, with an overall score of 96.8%. Few 2-site solutions are feasible, with the best being sites 3 and 5, yielding an overall score of 87.6%.

10.5. EXTENTIONS TO THE SELECTION MODEL

The previous sections have presented a basic prioritization model, (10.7) + $\mathrm{AR}(u_j, v_i)$. In the present section, some useful extensions will be sketched out in the context of prioritizing highway safety retrofit projects. We point out, however, that these extensions can have analogous interpretations in other problem settings such as the site selection example. In this context, the projects comprise specific sections of highway that are being considered for improvement from the viewpoint of accident potential or hazard. It is advantageous to use this alternative application setting, which itself instigated our work, as this will not only lend some perspective on the breadth of applicability of our basic model, but will also serve as an example of the necessity for, and ease of, its extension. These extensions will relax the assumptions concerning the independence of the projects and in so doing will exploit the binary structure in (10.7) to model details such as mutual exclusion/inclusion within subsets of projects.

Identifying hazardous sections of highway and prioritizing measures to improve them, in terms of reducing potential accidents, is a major consideration in all highway departments. A significant literature exists on the characterization of hazardous locations. The subject of interest here consists of two inter-related aspects. Firstly, there is a concern with the prediction of accident rates and their severity in terms of explanatory factors such as traffic levels, road geometrics and so on. This research has focused on the use of multiple regression as a mechanism for obtaining appropriate predictions (Head 1959). A second component of research in the road safety and accident analysis arena involves accident reduction factors; specifically the improvement in safety that will be achieved if a segment of highway network is modified in some way . See, for example, Persaud et al. (1992). This corpus of work, together with appropriate expert judgment, enables the prediction of benefits achievable consequent on the allocation of retrofit funds to specific project proposals.

With regard to prioritizing identified hazardous locations for treatment, the practice in most jurisdictions has been to rank these locations by either total accident frequency (e.g. using the total number of accidents on the road section over the past 3 years), or accident rate (e.g. accidents per million vehicle kilometers). A number of jurisdictions have recognized the multi-criteria nature of the prioritization problem. Thus, in looking at accident reductions, total accidents should properly be broken down into different severity classes such as fatal, major injury, minor injury, and property damage. In Kentucky, for instance, numerical weights are applied to various accident types to reflect costs to the public. Troxel (1993) discusses a number of severity models for combining fatal and injury accidents into an overall figure. There are, of course, different views in different jurisdictions of what the weights on different accident types should be. Persaud et al.(1992) present a multi-criteria methodology for determining appropriate weights to attach to different classes of accident in evaluating the relative importance of a set of retrofit measures. Further to these considerations, benefits can go beyond accident reduction and may also include improved road serviceability and traffic flow.

Thus, on the benefit (or output) side, the prioritization problem is clearly multi-dimensional. On the cost (or input) side, the multi-dimensional nature is also apparent. Obviously, the monetary expenditure required to implement a particular safety improvement, vis a vis the overall retrofit budget for the planning period, is the primary input at issue. Other factors may also impinge: availability of labour, plant, materials, and design office time, for example.

In this multi-output/multi-input setting, a model of the form dealt with in the previous sections, i.e (10.7) + $\mathrm{AR}(u_j, v_i)$, clearly has potential as a tool for assisting in the selection of a subset of projects from a larger collection of proposals. However, hitherto, we have assumed that projects are essentially independent, an assumption which now must be modified.

While in many prioritization settings the issue to be addressed is whether or not to undertake a given project, in the highway safety project prioritization problem there is at least one additional dimension; namely, the treatment or design choice. Specifically, there can be alternative ways to take corrective action at a particular hazardous site. For example, *run off road* accidents may be preventable or can be reduced either through shoulder upgrading (paving or widening), installation of guard rails, or even corrections to the geometry of the roadway. Each option has different associated outputs in terms of reductions in the various accident types and roadway serviceability as well as different calls on resources. Thus, there is a mutually exclusive set of treatments that may be applied for each hazardous site being considered. Model (10.7) can be easily modified to cater for this situation. Firstly, denote the set of distinct hazardous sites under consideration as Q and the set of (mutually exclusive) treatments being considered as T. The index set of all project variants under consideration thus becomes the cartesian product of sets Q and T:

$$P = Q \times T.$$

Subscripts referring to this index set, such as k and p, are now ordered pairs $\langle q,t \rangle$ where $q \in \mathrm{Q}$ and $t \in \mathrm{T}$. To ensure that no more than one project variant at a site q is selected, the following constraints are needed:

$$\sum_{t\in \mathrm{T}} c_{\langle q,t\rangle} \leq 1 \text{ for } q \in \mathrm{Q}$$

A second practical modeling requirement is that of specifying *commonality of treatment*, whereby a group of potential project sites in some geographic area should receive the same treatment. If, for example, shoulder widening is applied in a particular location to prevent accidents, it would normally be the case that this treatment would be implemented throughout the surrounding area. Thus, if A (subset of Q) is a set of sites to be so considered, we proceed as follows:

1. define binary variables $g_{\mathrm{A}t}$ for $t \in \mathrm{T}$,
2. include the constraint $\sum_{t\in \mathrm{T}} g_{\mathrm{A}t} \leq 1$
3. include the constraints

$$c_{\langle a,t\rangle} = g_{\mathrm{A}t} \qquad a \in \mathrm{A},\ t \in \mathrm{T}$$

or

$$c_{\langle a,t\rangle} \leq g_{\mathrm{A}t} \qquad a \in \mathrm{A},\ t \in \mathrm{T}$$

The first set of constraints in 3 implies that all (or none) of the sites in A be selected and treated identically, whereas the second version allows some of the sites to remain untreated.

10.6. CONCLUSING COMMENTS

In this chapter we have introduced the concept of selection within the DEA framework. This binary choice version of the DEA methodology opens a number of avenues for identifying specific groups of DMUs with certain desirable properties. An example of the use of this methodology for identifying core business components in a multi-plant firm environment is given in Cook and Green (2004), and as discussed in Chapter 11. Further research in this area is encouraged.

REFERENCES

1. Athanassopoulos, A.D. and J.E. Storbeck (1995), Non-parametric models for spatial efficiency, *Journal of Productivity Analysis* 6, 225-245.
2. Balakrishnan, A., A. Desai, and J.E. Storbeck (1990), Efficiency evaluation of retail outlet networks, *Environment and Planning B* 15, 131-141.

3. Banker, R.D., A.Charnes and W.W.Cooper (1984), Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science* 30, 1078-1092.
4. Bunch, D.S., D.Cardús, M.J.Fuhrer and R.M.Thrall (1989), When are additive models valid for evaluating proposed research, 28, 168-177.
5. Calvo, A. and D. Marks (1973), Location of health care facilities: An analytic approach. *Journal of Socio Economic Planning Sciences* 7, 407-484.
6. Charnes, A., W.W. Cooper, B. Golany, L.M. Seiford, and J. Stutz (1985), Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. *Journal of Econometrics* 30, 91-107.
7. Charnes, A., W.W.Cooper and E.Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2, 429-444.
8. Charnes, A., W.W.Cooper and R.M.Thrall (1991), A structure for classifying and characterizing efficiencies and inefficiencies in DEA, *Journal of Productivity Analysis* 2, 191-237.
9. Cook, W.D., and R. Green (2000), Project prioritization: A resource constrained data envelopment analysis approach, *Journal of Socio Economic Planning Sciences* 34, 85-99.
10. Cook, W.D., and R.H. Green (2003), Selecting sites for new facilities using data envelopment analysis, *Journal of Productivity Analysis* 19, 77-91.
11. Desai, A. and J.E. Storbeck (1990), Characterization of the constraints in successively-inclusive location hierarchies, *Environment and Planning B* 15, 131-141.
12. Eaton, D.J., R. Church, V. Bennett, B. Harmon, and L. Lopez (1982), On deployment of health resources in rural Valle de Cauca, Colombia. *Planning Processes in Developing Countries: Techniques and Achievements*, 17:331-359. Editors W.D. Cook and T.E. Kuhn.
13. Fisher, H., and G. Rushton (1979), Spatial efficiency of service locations and the regional development process, *Papers of the Regional Science Association*, 42:83-97.
14. Grosskopf, S. (1986), The role of reference technology in measuring productive efficiency, *Economic Journal* 96, 499-513.
15. Halmos, P.R. (1960), Naïve Set Theory, Van Nostrand Reinhold, New York.
16. Head, J.A. (1959), Predicting traffic accidents for highway elements on urban extensions of state highways, *Highway Research Board Bulletin*, 208, 45-63, Transportation Research Board, Washington, D.C.

17. Kao, C. (1998), Measuring the efficiency of forest districts with multiple working circles, *Journal of the Operational Research Society* 49, 583-590.
18. Maindiratta, A. (1990), Largest size-efficient scale and size efficiencies in data envelopment analysis, *Journal of Econometrics* 46, 57-72.
19. Martino, J.P. (1995), R&D Project Selection, Wiley, New York.
20. Oral, M., O.Kettani and P.Lang (1991), A methodology for collective evaluation and selection of industrial R&D projects, *Management Science* 37, 871-883.
21. Persaud, B.N., A.Kazakov and W.Cook (1992), Safety Expenditure and Achievable Benefits, *MTO Report ECN-92-01*, Research and Development Branch, Ministry of Transportation of Ontario, Toronto, Canada.
22. Savas, E. (1969), Simulation and cost-effectiveness analysis of New York's emergency ambulance service, *Management Science* 15, 608-627.
23. Thompson, R.G., F.D.Singleton Jr., R.M.Thrall and B.A.Smith (1986), Comparative site evaluations for locating a high-energy physics lab in Texas, *Interfaces* 16, 35-49.
24. Troxel, I. (1993), Severity models for roadside objects, *Transportation Research Circular 416*, Transportation Research Board/National Research Council, Washington, D.C.

This chapter is based upon Cook, W.D., and R.H. Green, 2000, Project prioritization: a resource-constrained data envelopment analysis approach, *Socio-Economic Planning Sciences*, 34:85-99, with permission from Pergamon Publishes, and Cook, W.D. and R.H. Green, 2003, Selecting sites for new facilities using data envelopment analysis, *Journal of Productivity Analysis*, 19:77-91, with permission from Kluwer Academic Publishers.

Chapter 11

MULTICOMPONENT EFFICIENCY

Measurement and Core Business Identification in Multiplant Firms

11.1. INTRODUCTION

The DEA model, developed by Charnes, Cooper and Rhodes (1978), provides a constant return to scale (CRS) methodology for evaluating the performance of a set of comparable decision making units (DMUs). In the usual setting, each DMU is evaluated in terms of a set of outputs that represent its accomplishments, and a set of inputs that represent the resources or circumstances at its disposal.

In some application areas, it has been recognized that the DMU may perform different *types* of functions. In such situations, it is desirable to derive a measure of performance, not only at the level of the DMU, but, as well, at the level of the particular function within the DMU. Cook and Roll (1993) were the first to examine the idea of partial efficiency measures, where the separate components of the DMU possess their own bundles of outputs and inputs. These bundles were assumed to be mutually exclusive of one another. Beasley (1995) examined both teaching and research components within a set of universities in the UK, and presented a nonlinear programming model for measuring DMU performance. A similar situation is encountered in Cook et al. (2001; 2000), where sales and service components are evaluated within a set of bank branches. They discuss linear models for providing both overall performance of a branch, as well as separate component performance measures. In that context, as with Beasley (1995) the input is a shared resource to be allocated to two production units. The complicating feature in each of these problem settings, that was not present in Cook and Roll (1993), is the presence of *shared resources*. The existence of *shared resources* means that the usual DEA structure must be modified to provide for a splitting of those resources among the various components.

In the current chapter we examine a set of manufacturing plants operating under a single corporate umbrella, with the objective of identifying how well each plant performs in each of its components thus identifying what might be considered each plant's a core business. Here, each component consists of a group of products selected from the totality of products offered, according to the specific interests of the corporate decision maker. Unlike the aforementioned dual-component applications (e.g., sales and service components in a bank branch), these components may overlap. Examples are (1) those products made from rolled steel of given dimensions; (2) those products servicing the automotive industry, ..., etc. This setting is clearly similar to those discussed above in that product groups are functions of the business, and, as will be seen, there are resources that are shared among those components. The models proposed here represent a departure from the earlier work of Beasley (1995) and Cook et al.(1993; 2001; 2002), in two respects. First, we examine the extension of the earlier models to a multi-component (two or more) setting. Second, using this multi-component structure as a point of departure, we develop models for *identifying the most appropriate* product groupings for each plant (DMU).

Section 11.2 presents the problem setting in more detail. In Section 11.3, extensions of the models of Cook et al. (2000; 2001) and Beasley (1995) are presented. Multiple, and potentially overlapping components are considered. These models are appropriate where the issue is one of identifying overall performance, as well as isolating particular areas (components) where the plant can be improved. Section 11.4 extends this idea to those situations wherein the organization wishes to identify the segment of the business that is performing *best* in any given DMU. In this way, the *core business* of each plant can be isolated, thus aiding the company in any reorganization initiatives designed to capitalize on the strengths of each location. Section 11.5 discusses the application of these models in the plant setting described earlier. Conclusions are given in Section 11.6.

11.2. MULTICOMPONENT EFFICIENCY MEASUREMENT AND CORE BUSINESS IDENTIFICATION

In this chapter we examine multi-component efficiency measurement from two perspectives. In the *first* situation, we make the assumption that the purpose of the performance assessment exercise is to determine an aggregate measure of efficiency, as well as measures for each of the separate components. Such evaluation will aid management in identifying the extent to which overall performance can be improved. As well, for specific business areas, the measures can point to those that are doing well, as well as those that require attention. Section 11.3 addresses this setting.

In the *second* situation, it is assumed that the organization wishes to go beyond simply identifying the level of performance of specific subunits of the business. Rather, it is desirable to identify the area(s) where DMUs are performing best, hence defining what might reasonably be regarded as each DMU's *core business.* A given DMU may then wish to focus its energies on this selected part of the operation, while de-emphasizing, or in some cases, even abandoning those portions of the business where it performs at a less than satisfactory level. This development is undertaken in Section 11.4.

To illustrate these ideas we examine a company with several plants that operate in the rolled steel industry. The company manufactures steel products, both of the *finished* variety that are sold on the open market, and semi-finished items that are custom-ordered, and sold to other manufacturers. These latter products can, for example, be items such as slit steel, used by other firms that manufacture steel doors and door frames. Other products, such as cylindrical bearings, are further along the value chain, and are purchased by companies that manufacture such consumer products as lawn mowers, or outboard motors for boats. Anticipating the detail given in Section 11.5, it is convenient to view the company's operations in terms of nine distinct products, and in conventional DEA terms each of these products would be considered an output. However, corporate management as well as being interested in the overall efficiency of each plant, is also interested in performance with respect to four overlapping groupings of these nine products. In what follows we will refer to a defined group of products, variously and interchangeably, as a component, subunit or segment. In some cases products are grouped to represent a particular market segment, e.g., automotive manufacturers who source certain products from the company. In other cases they are grouped to represent an internally meaningful segment of the operation, e.g., all products both semi-finished and finished, but pertaining to a certain size or quality of steel, or products made on particular machines.

In the section to follow, we present model structures for evaluating both the aggregate performance of each of a set of DMUs, as well as the performance of the separate subunits or components within a DMU's operation. For purposes of this development, we utilize the problem setting discussed herein as a backdrop.

11.3. MULTICOMPONENT MODEL STRUCTURES

The conventional model structure for evaluating the relative efficiency of each member of a set of DMUs is the DEA model of Charnes et al. (1978). Specifically, given an output vector $Y_k = (y_{1k}, y_{2k} \ldots, y_{Rk})$, and input vector $X_k = (x_{1k}, x_{2k}, \ldots, x_{Ik})$, for each of a set of n DMUs $k = 1, \ldots, n$, the constant returns to scale model is given by

$$\begin{aligned} &\max \ \mu_o Y_o / v_o X_o, \\ &\text{subject to} \\ &\mu_o Y_k / v_o X_k \leq 1, \quad \text{all } k, \\ &\mu_o, v_o \geq \varepsilon. \end{aligned} \tag{11.1}$$

The structure in (11.1) presumes that one desires to measure the overall efficiency (e.g., operational efficiency) of each DMU, without consideration for the performance of subunits that may exist within the DMU. In the problem setting presented herein, we wish to provide for a more detailed performance evaluation, i.e., at the level of these subunits.

11.3.1 Multi-component Efficiency Measurement with Shared Inputs: Non-overlapping Subunits

Our point of departure for the discussion in this section, is the model structures of Cook et al. (2000), (see also Cook and Hababou (2001)). There, the authors examine the problem of providing separate efficiency measures for both *sales* and *service* components of a set of bank branches for a major Canadian bank. Adopting the notation of Cook et al. (2000), and extending their model structure to "T" components, we have:

Parameters:

Y_k^t = the R-dimensional vector of outputs included in the t th component of DMU k.
R = set of all outputs
R^t = set of outputs generated by the t th component.
X_k^t = the I-dimensional vector of inputs dedicated to the t th component of DMU k.
I = set of all inputs.
I^t = set of inputs dedicated to the t th component.
X_k^S = the I^s-dimensional vector of inputs shared among the T components of DMU k.
I^s = set of shared inputs.
L_i^t, U_i^t = lower, upper limits on the portion of the i th shared resource, that can be assigned to the t th component of a DMU.
T = set of all components.

Decision Variables:

μ_p^t = vector of multipliers applied to outputs Y_q^t.
$\nu_{q_t}^t$ = vector of multipliers applied to inputs X_o^t.
ν_o^{st} = vector of multipliers applied to that portion of shared inputs X_o^S that are assigned to component t.
α_o^t = vector representing the proportion of shared inputs X_o^S allocated to the t th component.

In the two-component problem addressed in Cook et al. (2000), the principal area of difficulty was the presence of shared inputs X_k^S. Specifically, there are certain resources such as branch expenditure on computer technology and general branch staff, that are shared across the two components of the business. There is no well defined split of these resources across different functions, and the basic problem has to do with the allocation of these inputs among the components. To facilitate this, and at the same time extend the idea to the general case of T components, a decision vector α_k^t is introduced that permits the DMU k in question to apportion X_k^S among the T competing components. In Cook et al. (2000), this is done in a manner that optimizes the *aggregate* performance measure (of DMU "o") given by:

$$e_o^a = \sum_{t\varepsilon T} \mu_o^t Y_o^t / [\sum_{t\varepsilon T} (\nu_o^t X_o^t + \nu_o^{st}(\alpha_o^t X_o^S))] \quad (11.2)$$

The component-specific performance measures e_o^t are given by:

$$e_o^t = \mu_o^t Y_o^t / (\nu_o^t X_o^t + \nu_o^{st}(\alpha_o^t X_o^S)) \quad (11.3)$$

It is pointed out that the notation $\alpha_o^t X_o^S$ represents the vector $(\alpha_{o1}^t x_{o1}^S, \alpha_{o2}^t x_{o2}^S, \ldots, \alpha_{oI^S}^t x_{oI^S}^S)$ of shared inputs allocated to component t by DMU "o".

In the discussion below, we distinguish between optimal performance measures and performance measures for a DMU k, evaluated in terms of the multipliers for a DMU "o" currently being considered. (Doyle and Green (1994) use the term *cross-evaluation* in this instance). For this purpose, we adopt the notation $\hat{e}_k^a, \hat{e}_k^t$ to denote the measures for DMU k that represent their optimal performance, while e_k^a, e_k^t denote performance relative to multipliers arising from the optimization of (some other) DMU "o".

The multi-component DEA model is given by:

$$\hat{e}_o^a = \max e_o^a \quad (11.4a)$$

subject to

$$e_k^t \le 1, \text{ all } t,k \quad (11.4b)$$

$$L_i^t \le \alpha_{oi}^t \le U_i^t \text{ all } t,\ i\ \varepsilon\ I^S, \quad (11.4c)$$

$$\sum_{t\varepsilon T} \alpha_{oi}^t = 1,\ i\ \varepsilon I^S, \qquad (11.4d)$$

$$\mu_o^t, \nu_o^t, \nu_o^{st} \geq \varepsilon, \text{ all } t. \qquad (11.4e)$$

Here, the objective (11.4a) maximizes the overall performance measure for the DMU "o", in the spirit of the original DEA model of Charnes et al. (1978). Correspondingly, we restrict each component measure e_k^t by an upper bound of 1 in (11.4b). A permissible range on the proportion of the ith shared resource that can be allocated to the tth component by any DMU is given by (11.4c). Constraints (11.4d) specify that the proportional splits of any input i across the T components sum to unity. Finally, constraints (11.4e) restrict multipliers to be strictly greater than zero.

The limits L_i^t, U_i^t, on the proportions α_{oi}^t of the various inputs i to components t would need to be specified by the user. Such limits might generally arise from any information available at the plants regarding standard amounts of inputs i per unit of product in components t.

From the above discussion it is clear that problem (11.4) is a *restricted version* of problem (11.1). Specifically, any feasible solution to (11.4) is also feasible for (11.1). Problem (11.4) only permits multipliers which identify each component of the plant as a bona fide sub-DMU whose performance measure is captured at the same time as that of the entire plant. Problem (11.1), however, is focused purely at the plant level, with no recognition whatever of subunits.

Definition 11.1: A DMU "o" is said to be *efficient* if its aggregate score $\hat{e}_o^a = 1$.

Definition 11.2: A DMU "o" is said to be *efficient in its t th component* if $\hat{e}_o^t = 1$.

Theorem 11.1: In model (11.4), the resulting aggregate performance measure $\hat{e}_k^a$ for any DMU k, does not exceed unity, i.e., $\hat{e}_k^a \leq 1$.

Proof:

If we define

$$\beta_k^t = [\nu_o^t X_k^t + \nu_o^{st}(\alpha_o^t X_k^s)] / \sum_{t\varepsilon T} (\nu_o^t X_k^t + \nu_o^{st}(\alpha_o^t X_k^s)),$$

then, the aggregate measure (in terms of the (μ_o, ν_o) multipliers), is given by

$$\hat{e}_k^a = \sum_{t\varepsilon T} \beta_k^t \hat{e}_k^t.$$

Hence, e_k^a is a convex combination of the component measures, and as such $e_k^a \leq 1$. Q.E.D

Theorem 11.2: In model (11.4), a DMU is efficient if and only if it is efficient in each of its components.

Proof:

Case 1: Assume all component measures $\hat{e}_k^t = 1$.

By definition,

$$\hat{e}_k^a = \sum_{t \varepsilon T} \beta_k^t \hat{e}_k^t$$

from Theorem 11.1, and since $\sum_t \beta_k^t = 1$, it follows that $\hat{e}_k^a = 1$.

Case 2: Assume $\hat{e}_k^a = 1$. Then, if any $\hat{e}_k^t < 1$, it must be the case that

$$\hat{e}_k^a = \sum_{t \varepsilon T} \beta_k^t \hat{e}_k^t < 1,$$

as well, in contradiction. Q.E.D.

We now examine multi-component performance measurement when overlaps can occur.

11.3.2 Multi-component Efficiency Measurement with Overlapping Subunits

The models presented above presume a set of subunits that are mutually exclusive. Arguably, in the bank branch setting of Cook and Hababou (2001), and Cook et al. (2000), sales and service components meet the mutual exclusivity requirement. In many settings this restriction may not hold, however, as is the case with the business components described later.

In the case where mutual exclusivity prevails, it is sufficient to subdivide a shared input among the set of components. That is, α_{oi}^t represents the portion of input i assigned to component t. It is not necessary to address how this portion α_{oi}^t is distributed among the outputs comprising component t. In case there is *overlap* among the components due to the existence of common outputs, the manner in which the proportions $\{\alpha_{oi}^t\}_{t=1}^T$ behave, is no longer clear. It is, for example, not true that $\sum_{t \varepsilon T} \alpha_{oi}^t = 1$, due to the overlap.

In recognition of the *overlap problem*, we need to be more exacting as to how the shared input i is assigned to outputs $r \varepsilon \Re$. Specifically, we define variables α_{oir} that denote the proportion of shared input x_{oi}^s (the ith component of vector X_o^s) that is allocated to output y_{or}. As well, let L_i^r, U_i^r, denote lower and upper bounds, respectively, on α_{oir}, and impose the constraint

$$\sum_{r \in \Re} \alpha_{oir} = 1.$$

The proportion α_{oi}^t of input i allocated to component t is then the sum of the proportions α_{oir} of i allocated to those outputs comprising t, i.e.

$$\alpha_{oi}^{t}=\sum_{t\varepsilon\Re^{t}}\alpha_{oir}.$$

For brevity in modelling, we henceforth denote the feasible set of $\alpha=(\alpha^{t})$ by

$$\Lambda_{o}=\{\alpha_{o}=(\alpha_{o}^{t}):(1)\,\alpha_{oi}^{t}=\sum_{r\varepsilon\Re^{t}}\alpha_{oir};$$
$$(2)\ L_{i}^{r}\le\alpha_{oir}\le U_{i}^{r};(3)\sum_{r\varepsilon\Re}\alpha_{oir}=1,$$
$$\alpha_{oir}\ge 0,\text{all } i\,\varepsilon\,\mathrm{I}^{s},\text{ all } t\}.$$

The multi-component DEA model is then given by:

$$\text{Max } e_{o}^{a}, \tag{11.5a}$$
subject to
$$e_{k}^{t}\le 1,\text{ all } t,k, \tag{11.5b}$$
$$\alpha_{o}\varepsilon\,\Lambda_{o}, \tag{11.5c}$$
$$\mu_{o}^{t},\ v_{o}^{t},\ v_{o}^{st}\ge\varepsilon,\ \text{ all } t. \tag{11.5d}$$

It is noted that the objective function (11.5a) credits the DMU for producing an output y_{or}^{t} as many times as that output appears as a member of a component's output set. For example, an output y_{or}, contained in both components $t=1$ and $t=2$, (i.e., $y_{or_1}^{1}=y_{or_2}^{2}$), would appear in (11.5a) twice, as $\mu_{or_1}^{1}y_{or_1}$ and $\mu_{or_2}^{2}y_{or_2}$.

We point out, however, that, as in the case of non-overlapping subunits, it is also true here that problem (11.5) is simply a restricted version of problem (11.1), if we view the inputs X in (11.1) as all being shared inputs. This is captured by the following theorem.

Theorem 11.3:
Any feasible solution to problem (11.5) is feasible to (11.1).

Proof: Define the R-dimensional multiplier vector $U^{t}=(u_{r}^{t})$ by

$$u_{r}^{t}=\begin{cases}\mu_{r}^{t} & \text{if product r is in component t}\\ 0 & \text{otherwise}\end{cases}$$

and let $U=\sum_{t\varepsilon T}U^{t}$. Letting Y denote the R-dimensional vector of all outputs as used in (11.1), it follows that

$$\sum_{t\varepsilon T} \mu^t Y^t = UY.$$

Similarly, one can replace the set of inputs $\{X^t\}$ by the I-dimensional vector $X(1) = (X^1, X^2, ..., X^T)$, and replace the set of "shared resource" vectors $\alpha^t X^s$ by the sum of these component shares to get X^s. Let $X = (X(1), X^s)$, the full vector of all inputs. Then, as with the output side, one can express the denominator of the performance measure as

$$\sum_{t\varepsilon T} [\nu^t X^t + \nu^{st}(\alpha^t X^s)] = VX,$$

where V is defined in terms of the ν^t, ν^{st} in a manner analogous to the definition of U in terms of $\{\mu^t\}$. Hence e^a_o in (11.5) can be written as

$$e^a_o = UY/VX.$$

Since it is true that each component measure $e^t_k \leq 1$, then it must also be true that the aggregate score $e^a_k \leq 1$ as well. Thus, any feasible solution to (11.5) is also feasible for (11.1).

Q.E.D.

Hence, the overlap of the components does not lead to inconsistencies in regard to problem (11.1). Defining the aggregate measure in this manner results in the following theorem. The Proof is analogous to those of Theorems 11.1 and 11.2, and is, therefore, omitted.

Theorem 11.4

(a) The aggregate measure of efficiency given by (11.5a) does not exceed unity.

(b) A DMU will be aggregate-efficient, (the objective function (11.5a) will equal unity), if and only if it is efficient in each component measure.

Model (11.5), thus, allows one to examine the performance of a DMU in each business area. As well, it provides an overall or aggregate measure of performance across all business components.

Because the orientation of model (11.5) is toward evaluation of the DMU at an aggregate level, with component measures arising only as a by-product, it can be argued that the individual subunits of the business may not be shown in their most favorable light. In some cases, the strategic intent of the organization might be to identify the core business for each DMU, the purpose being to focus the attention of the DMU toward the areas of the business at which it performs best. In the section to follow, we present model structures wherein the intention is to choose a core business component on behalf of each DMU. It should be pointed out that the identification of a core business component will not necessarily imply the immediate termination of all activities at a plant that are not included in that component. Rather, a DMU would initially continue to service all existing activities, possibly phasing out non-core activities as these are redistributed to where they are best accomplished over some time horizon.

11.4. MODELLING SELECTION OF CORE BUSINESS COMPONENTS

A typical problem setting would be one where each of a set of plants for a given company produces a full product line, for sale and distribution to customers. There can be a number of reasons why it is cost effective for a certain product line, for example, to be manufactured in particular locations, but not in others. Certain manufactured items may, for instance, require specialized and expensive equipment that the company might prefer to make available in only one location. Alternatively, certain customers (e.g. farmers) may be highly concentrated in one geographical area, meaning that a plant close to that concentration should produce products related to that customer group. As well, simple economies of scale may dictate that the production for a product be concentrated in only a few plants, or even a single plant.

The problem then is to identify which collection of products or product lines should be handled by any given plant, thus defining that plant's core business.

The conventional DEA model does not readily lend itself to resource allocation (i.e. allocation of shared inputs). The DEA approach focuses attention on the performance of a particular DMU "o". If the objective is to allocate components to DMUs (plants), and to divide shared resources among products (and thus among components), one needs to view this allocation process from the perspective of the entire *collection* of DMUs, simultaneously rather than from the conventional DEA perspective, i.e. iteratively, one DMU at a time.

To facilitate the allocation of components to DMUs, define the bivalent variables $\{d_k^t\}_{t=1}^T$, for each DMU k,

$$d_k^t = \begin{cases} 1 & \text{if component } t \text{ is assigned to DMU } k, \\ 0 & \text{otherwise.} \end{cases}$$

The aggregate performance (ratio) measure for the collection of DMUs, given an allocation defined by a chosen set of d_k^t values, can be expressed as:

$$\frac{\sum_k [\sum_t d_k^t \mu^t Y_k^t]}{\sum_k [\sum_t d_k^t (v^t X_k^t + v^{st}(\alpha_k^t X_k^s))]}$$

The *optimal* assignment of components to DMUs, as defined by the d_k^t, is arguably that for which the ratio of aggregate output to aggregate input is maximized. The set of d_k^t for which this maximum occurs can be determined by solving the fractional programming problem:

$$\max \sum_k [\sum_t d_k^t \mu^t Y_k^t] / \sum_k [\sum_t d_k^t (v^t X_k^t + v^{st} (\alpha_k^t X_k^s))] \quad (11.6a)$$

subject to

$$\mu^t Y_k^t / (v^t X_k^t + v^{st} (\alpha_k^t X_k^s)) \leq 1, \text{ all } k,t \quad (11.6b)$$

$$\alpha \, \varepsilon \, \Lambda_0, \quad (11.6c)$$

$$\sum_t d_k^t \geq 1, \text{ all } k, \quad (11.6d)$$

$$\sum_k d_k^t \geq 1, \text{ all } t, \quad (11.6e)$$

$$\mu^t, v^t, v^{st} \geq \varepsilon, \text{ all } t, \ \ d_k^t \, \varepsilon \{0,1\}, \text{ all } k,t. \quad (11.6f)$$

Constraints (11.6b) restrict the ratio of outputs to inputs in *any* component to not exceed unity. (11.6c) requires that the resource splitting variables satisfy conditions as defined earlier in Λ_o. Constraints (11.6d) force each plant k to support *at least* one product group or component. Similarly, (11.6e) stipulates that each component must be produced at *one or more* of the plants.

It is conceivable that at the optimum, certain plants may be chosen to support several product groups, while other plants may service only one group.

Model (11.6a)-(11.6f), assigns multipliers μ^t, v^t, v^{st} to each component t in each DMU k. While it is not the purpose of the model to measure the efficiency of the *entire* operation of each plant, the supplied (common set of) multipliers do in fact provide the basis for an efficiency score for each plant and the aggregate across all plants, should one want to extract these. That aggregate score clearly includes the contribution rendered by both core and non-core components of the plant. Admittedly, the set of multipliers is derived in a manner designed to display *core* components in their best light, and by implication, non-core components in a light less than best. Hence, non-core components may be represented in a disadvantageous manner. One might argue that this is appropriate since, over time, such non-core components will, in any event, be phased out. Thus, their estimated performance (by that stage) will be a non-issue. At the same time, the model does, in fact, recognize their existence, and the bounds $[L_i^r, U_i^r]$ appropriately force the allocation of shared resources across all components (both core and non-core). Thus, choice of these bounds by management affirms the continuing presence of non-core components in the operation.

Thus, the real purpose of the model is to single out those components of each plant on which that plant exhibits its best performance. It is these core components whose aggregate performance we wish to capture.

The implication of this is that when a set of plants exhibit inefficiency, it is often desirable to strive for *specialization*. The questions that management would like to answer are:

(1) In what parts of the operation should each plant specialize?
(2) If plant operations were reorganized to implement such specialization, what would be the anticipated performance of the resulting operation?
(3) How would each reorganized (future) plant perform?

Question 1: The purpose of the model is to extract those components at each plant that appear to be the ones in which the plant should specialize.

Question 2: While the model yields an aggregate performance across all core components in all plants, there is an implied measure of performance for each plant on a portion (core business portion) of that plant's operation. Specifically, using $\{\hat{d}_k^t\}_{t=1}^T$, for each k, the model yields a measure of performance for that subset of components in terms of the inputs that those components utilize, and the outputs generated by those components. This measure captures how the (reduced) plant *would* perform if non-core business elements were not present.

Question 3: In a reorganized structure, the essence of the model is that each plant would concentrate only on its core business activities. It is argued that if each plant were to scale up its operation such as to come to full capacity in its resource utilization, then it is hypothesized that the resulting output generated would be scaled up by the same factor.

To solve problem (11.6a)-(11.6f), it can be shown that it is representable as a mixed integer linear programming problem. This is given by the following theorem.

Theorem 11.5

Problem (11.6a)-(11.6f) can be represented as a mixed integer (binary) linear problem.

Proof: Problem (11.6a)-(11.6f) is equivalent to the mixed binary *nonlinear* programming model:

$$\max \sum_k \sum_t d_k^t \mu^t Y_k^t \tag{11.7a}$$

subject to

$$\sum_k \sum_t d_k^t (\nu^t X_k^t + \nu^{st}(\alpha_k^t X_k^s)) = 1 \tag{11.7b}$$

$$\mu^t Y_k^t - (\nu^t X_k^t + \nu^{st}(\alpha_k^t X_k^s)) \le 0, \text{ all } k, t, \tag{11.7c}$$

$$\alpha \,\varepsilon\, \Lambda_o, \tag{11.7d}$$

$$\sum_t d_k^t \ge 1, \text{ all } k, \tag{11.7e}$$

$$\sum_k d_k^t \ge 1, \text{ all } t, \tag{11.7f}$$

$$\mu^t, \nu^t, \nu^{st} \ge \in, \text{ all } t,\ d_k^t \,\varepsilon\, \{0,1\}, \text{ all } k, t. \tag{11.7g}$$

Make the change of variables:

$$\overline{\nu}^{st} = \nu^{st}\alpha^t,\ \mathrm{v}_k^{st} = d_k^t \overline{\nu}^{st},\ \mathrm{v}_k^t = d_k^t \nu^t,\ u_k^t = d_k^t \mu^t.$$

It is noted that we can replace an expression such as $v_k^t = d_k^t \nu^t$ by the constraint set

$$v_k^t \le M d_k^t,$$

$$\nu^t \ge v_k^t,$$

$$\nu^t \le v_k^t + M(1 - d_k^t),$$

where M is a large positive number. Specifically, if $d_k^t = 0$, then $v_k^t = 0$; if $d_k^t = 1$, then $v_k^t = \nu^t$. A similar set of constraints can be imposed to replace the nonlinear expressions $u_k^t = d_k^t \mu^t$, and $v_k^{st} = d_k^t \upsilon^{st}$. Problem (11.7a)-(11.7g) can then be written as the mixed binary linear programming model

$$
\begin{aligned}
&\max \sum_{k} \sum_{t \varepsilon T} u_k^t Y_k^t, \\
&\text{subject to} \\
&\sum_{k} \sum_{t \varepsilon T} (v_k^t X_k^t + v_k^{st} X_k^s) = 1, \\
&\sum_{t \varepsilon T} [u_k^t Y_k^t - (v_k^t X_k^t + v_k^{st} X_k^s)] \le 0, \text{ all } k, \\
&\upsilon_k^t \le \mathrm{M} d_k^t, \text{ all t}, \\
&\nu^t \ge \mathrm{v}_k^t, \text{ all t}, \\
&\nu^t \le \mathrm{v}_k^t + \mathrm{M}(1 - \mathrm{d}_k^t), \text{ all t}, \\
&u_k^t \le \mathrm{Md}_k^t, \text{ all t}, \\
&\mu^t \ge u_o^t, \text{ all t}, \\
&\mu^t \le u_k^t + \mathrm{M}(1 - \mathrm{d}_k^t), \text{ all t}, \\
&\upsilon_k^{st} \le \mathrm{Md}_k^t, \text{ all t}, \\
&\bar{\nu}^t \ge \mathrm{v}_k^{st}, \text{ all t}, \\
&\bar{\nu}^t \le \mathrm{v}_k^{st} + \mathrm{M}(1 - \mathrm{d}_k^t), \text{ all k,t}, \\
&\alpha \, \varepsilon \, \Lambda, \\
&\sum_{t} \mathrm{d}_k^t \ge 1, \text{ all k}, \\
&\sum_{k} \mathrm{d}_k^t \ge 1, \text{ all t}, \\
&\bar{\nu}_i^{st} \ge \varepsilon \alpha_i^t, \text{ all i,t}, \qquad (11.8) \\
&\mu_r^t, \nu_i^t \ge \varepsilon, \text{ all i,r,t}, \\
&u_{kr}^t, \mathrm{v}_{ki}^t \ge 0, \text{ all i,r,k}, \\
&\upsilon_{ki}^{st} \ge 0, \text{ all r,t,i} = 1,..,\mathrm{I}^S, \\
&\mathrm{d}_k^t \varepsilon \{0,1\}, \text{ all k,t}.
\end{aligned}
$$

This completes the proof.

QED

There are clearly variations of this model where, for example, it may be pertinent for certain product groupings or components to be manufactured in only certain plants that are perhaps in the best possible position to handle them. This might be due to equipment capability, proximity of the market, and so on. Thus, for a given component t_o, we might require that

$d_k^{to} = 0, k\varepsilon \overline{K}^{to}$, where K^{to} is the set of allowable plants for manufacturing component t_o, and $\overline{K}^{to}$ is its compliment.

In the section to follow, this model is used to allocate business components to ten plants within the company under study.

11.5. APPLICATION OF CORE BUSINESS SELECTION MODEL TO A SET OF PLANTS

In the problem studied, 10 plants currently operate under a single corporate umbrella, producing a variety of steel products including bearings, pipes and sheet steel of various sizes. Clearly, some of these products are of the finished goods variety (e.g. pipes), while others are semi-finished, becoming components in other manufactured items (bearings), or are sold to other plants for further manufacturing (sheet steel).

As indicated earlier, it is convenient to view each plant's business as consisting of various components. While it is the case that there can be a large number of products to consider (e.g. different sizes of circular bearings), here items have been grouped by management under a few major categories. For purposes of this study we present the operation of any plant as consisting of four (overlapping) components, defined by their outputs y_r^t, the number of units of output r in the t th component:

Component #1:

- All solid bearings (y_1^1)
- Circular bearings (automotive) (y_2^1)
- Sheet steel $\leq$ 4 feet in length (y_3^1).

Component #2:

- Solid bearings (automotive) $y_1^2)$
- Steel pipes $\leq$ 8 feet in length $y_2^2)$
- Sheet steel 4 feet to 8 feet in length (y_3^2)

Component #3:

- Steel pipes $>$ 8 feet in length (y_1^3)
- Sheet steel $>$ 8 feet in length (y_2^3)

Component #4:

- Circular bearings (automotive) (y_1^4)
- Circular bearing (non-auto) (y_2^4)

- All solid bearings (y_3^4)
- Sheet steel $\leq$ 4 feet in length (y_4^4)

Table 11-1 displays the data for all outputs for the 10 plants considered.

Table 11-1. Outputs for Four Components

Plant	y_1^1	y_2^1	y_3^1	y_1^2	y_2^2	y_3^2	y_1^3	y_2^3	y_1^4	y_2^4	y_3^4	y_4^4
1	50	30	70	30	60	50	40	80	30	50	50	70
2	45	35	60	25	50	50	40	75	35	55	45	60
3	75	25	50	35	55	40	50	70	25	60	75	50
4	60	40	80	40	40	30	70	50	40	50	60	80
5	35	25	25	20	25	20	35	20	25	30	35	25
6	55	60	40	40	60	45	60	50	60	50	55	40
7	120	100	100	100	80	120	120	60	100	110	120	60
8	60	80	25	50	100	20	80	35	80	80	60	25
9	25	75	65	20	25	80	100	70	75	70	25	65
10	100	55	40	70	35	65	35	45	55	60	100	40

The resources committed to the production of these product lines can be grouped under four headings, namely

- Shop labour (x_1)
- Machine labour (x_2)
- Steel splitting equipment (x_3)
- Lathes (x_4)

Shop labour and machine labour are measured in full time equivalent (FTE) staff. Both equipment variables are expressed in hundreds of hours of capacity available per month. Given the manner in which the four components have been defined, with the inherent overlap of products, all four of these inputs should be viewed as shared resources.

Table 11-2 shows the amounts of the four resources corresponding to each plant.

The connection between the shared inputs and the product outputs (y_r^l) is quite complex, and must be reflected through the α_{ir}. If a given input such as lathes (x_4) does not impact on a particular output such as sheet steel ($\leq$ 4feet) (y_3^1) then that particular variable α is set to zero. Figure 11-1 shows the input-to-output impact matrix.

In the figure, an "x" denotes the fact that the particular input contributes to the output shown. It must be noted as well, that when we have a product common to two or more components, the corresponding variables α_{ir} must be equated. For example, since sheet steel $\leq$ 4 feet is part of both components 1 and 4 (i.e., $y_3^1 = y_4^4$), then $\alpha_{1,3} = \alpha_{1,12}$.

Table 11-2. Shared Resources

DMU	x_1	x_2	x_3	x_4
1	30	15	100	150
2	40	12	90	180
3	35	16	97	100
4	38	20	85	85
5	28	9	110	125
6	37	13	76	140
7	31	18	83	110
8	35	15	100	150
9	25	19	95	190
10	30	10	65	210

Input	y_1^1	y_2^1	y_3^1	y_1^2	y_2^2	y_3^2	y_1^3	y_2^3	y_1^4	y_2^4	y_3^4	y_4^4
x_1	—	—	x	—	x	x	x	x	—	—	—	x
x_2	x	x	—	x	—	—	—	—	x	x	x	—
x_3	—	—	x	—	x	x	x	x	—	—	—	x
x_4	x	x	—	x	—	—	—	—	x	x	x	—

Figure 11-1. Input Versus Output Impact Matrix

For solution purposes we have restricted each α_{ir} to lie in the range .1 to .4. This means that for each shared input i, at least 10% and not more than 40% of that input would be dedicated to any given output r. Although the decision on such bounds was difficult for management to pin down, the .1-.4 range was deemed reasonable. As well, we impose both upper and lower limits on the numbers of plants to which any given component can be assigned. Specifically, we require for each component t :

$$1 \leq \sum_k d_k^t \leq 4.$$

Hence, at least one plant, and no more than four plants can be assigned component t.

Efficiency Results

Table 11-3 displays the optimal component assignment to plants. In summary:

Component #1 $\longrightarrow$ Plants 5,7,10
Component #2 $\longrightarrow$ Plants 6,8

Component #3 ⟶ Plants 1,3,9
Component #4 ⟶ Plants 2,4

Table 11-3. Assignment of Components to Plants

DMU	T_1	T_2	T_3	T_4
1	0	0	1	0
2	0	0	0	1
3	0	0	1	0
4	0	0	0	1
5	1	0	0	0
6	0	1	0	0
7	1	0	0	0
8	0	1	0	0
9	0	0	1	0
10	1	0	0	0

Table 11-4. Decomposition of DMU Efficiency

DMU	Assignment of components to plants				Partial efficiencies of component				Aggregate efficiency
	T_1	T_2	T_3	T_4	T_1	T_2	T_3	T_4	
1	0	0	1	0	0.51	0.68	1.00	1.00	0.93
2	0	0	0	1	0.54	0.68	1.00	1.00	0.94
3	0	0	1	0	0.52	0.61	0.90	0.76	0.75
4	0	0	0	1	0.53	0.39	0.76	1.00	0.86
5	1	0	0	0	0.43	0.45	0.25	0.59	0.47
6	0	1	0	0	0.60	0.79	0.83	0.79	0.78
7	1	0	0	0	1.00	1.00	1.00	1.00	1.00
8	0	1	0	0	0.56	1.00	0.50	0.56	0.60
9	0	0	1	0	0.45	0.34	1.00	0.84	0.78
10	1	0	0	0	0.81	0.78	0.85	0.86	0.85

The overall efficiency score corresponding to this assignment is 96.6% (the value of objective function (11.7a)). Specifically, if plants are evaluated only on their core business components, their performance will be such that if viewed as a single entity, the aggregate score is 96.6%. Table 11-4 displays both the current aggregate efficiencies for the 10 plants, as well as a decomposition of these scores into component efficiencies. For example, Plant #3 currently displays an overall performance score of 75%. This is composed of partial efficiency scores of 52%, 61%, 90% and 76% for components 1, 2, 3 and 4, respectively. Recall that the measure of partial efficiency for a DMU k in its t th component is given by

$$e_k^t = \mu^t Y_k^t / [v^t X_k^t + \alpha_k^t v^{st} X_k^S].$$

It is noted, as well, that with the recommended component-to-plant assignments, plant #3 would be expected to have an efficiency of 90% (up from 75%), if it could ultimately phase out non-productive portions of its operation, and move its full emphasis to that part of the business defined by component #3. It must be emphasized that the component t_0 assigned to a plant may not be the one whose partial efficiency is highest for that plant. Notice, for example, that component #2 is assigned to plant #6, with a partial efficiency of 79%, yet component #3 actually performs better within that plant (at a partial efficiency of 83%). This can occur because rather than minimizing a sum of efficiency ratios, we are optimizing the ratio of aggregate output (across all plants), to aggregate input.

11.6. DISCUSSION

This chapter has examined the problem of identifying core business components for each of a set of comparable decision making units. In the context of a set of manufacturing plants, a modified version of the DEA model of Charnes et al. (1978) has been developed and demonstrated. Unlike conventional applications of DEA where the scope of the business (bundle of products produced) is assumed to remain fixed, the approach herein is intended to aid in making decisions pertaining to functional specialization in plants. An important by-product of the core-business selection process is the evaluation of efficiency of each component of the business as well as of the overall DMU. The result, as demonstrated by Table 11-4, is an efficiency profile that management can utilize in deciding where to aim for improvements and, as well, which components to de-emphasize or phase out.

We do not attempt to address issues relating to plant reorganization toward specialization. Rather, the model can aid management in choosing those (core) business activities to place within each plant. The logistics of restructuring and any change management considerations are beyond the scope of the current chapter.

One of the potential shortcomings of the model given here is the apparent absence of consideration of distribution costs on the input side. Specifically, in some settings, the choice of a particular plant as the location out of which a given component of the business will be operated, has distributional consequences. For example, manufacturing auto parts in a location remote from automobile plants (the customer) may be more costly than having them manufactured at a less efficient, but closer-to-market facility. In the application discussed herein, this issue was not highlighted as a major concern. Presumably, in situations where distribution is a major issue, one would need to augment the input bundle to include a provision for distribution costs.

REFERENCES

1. Banker, R.D., A. Charnes, W.W. Cooper (1984), Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science* 30, 1078-1092.
2. Beasley, J.E. (1995), Determining teaching and research efficiencies, *Journal of the Operational Research Society* 46, 4451-452.
3. Charnes, A., W.W. Cooper, B. Golany, L.M. Seiford, J. Stutz (1985), Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions, *Journal of Econometrics* 30, 91-107.
4. Charnes, A., W.W. Cooper, E.L. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2, 429-444.
5. Cook, W.D., M. Hababou (2001), Sales performance measurement in bank branches, *OMEGA*, 29, 299-307.
6. Cook, W.D., M. Hababou, H.J.H. Tuenter (2000), Multi-component efficient measurement and shared inputs in data envelopment analysis: An application to sales and service performance in bank branches, *Journal of Productivity Analysis*, 14, 209-224.
7. Cook, W.D., Y. Roll (1993), Partial efficiencies in data envelopment analysis, *Socio-Economic Planning Sciences* 37, 171-179.
8. Cook, W.D., and R.H. Green (2004), Multi-component efficiency measurement and core business identification in multi-plant firms: A DEA model, *European Journal of Operational Research*, 157, 540-551.
9. Cooper, W., K. Park, J. Pastor (1999), RAM: A range adjusted measure of inefficiency for use with additive models, and relations to other models and measures in DEA, *Journal of Productivity Analysis* 11, 5-42.
10. Doyle, J.R., R.H. Green (1994), Efficiency and cross efficiency in DEA: Derivations, meanings and uses, *Journal of the Operational Research Society* 45, 567-578.

This chapter is based upon Cook, W.D. and R.H. Green, (2004), Multicomponent efficiency measurement and core business identification in multiplant firms: A DEA model, European Journal of Operational Research, 157, 540-551, with permission from Elsiever Publishing

Chapter 12

IMPLEMENTATION OF ROBOTICS

Identifying Efficient Implementors

12.1. INTRODUCTION

One significant component of the factory of the future is the industrial robot.[1] Statistics from various countries show significant growth in the number of robots and plants using robots since 1980. It is becoming increasingly important to understand which firms are doing an adequate job of implementing robotics, and why they have been successful. Determining factors that lead to better robotics implementation is complicated by the difficulty in determining project success. Performance is often defined in terms of multiple criteria, and the levels for each should be interpreted in relation to similar applications in firms with similar or competing interests. This presents a complex situation in which to identify the more or less successful implementers. It is only after the better performers are identified that one can weigh the merits of various robotics implementations.

This chapter introduces a model of implementation efficiency which utilizes DEA methodology to rank implementation performance in the presence of multiple criteria. The source data are from a field study of robotic applications in 30 companies. Three conditions believed to influence implementation efficiency are tested, illustrating how DEA ratings can be used to study the implementation of new technology.

[1]A robot is defined as a programmable multifunctional manipulator designed to move parts, tools or specialized devices through variable programmed motions for the performance of a variety of tasks.

12.2. A MODEL OF IMPLEMENTATION EFFICIENCY

12.2.1 Background

In the research on technological innovation, the primary focus has been on the decision phase, that is, the stages leading up to the organization's choice to adopt new technology. The later stages of the adoption process, during which the innovation is actually implemented, have generally received much less attention. There are a number of reasons for this:

- The implementation events are contingent upon a great many factors, including those related to:
  - the technology's characteristics, such as its physical complexity and its state of development;
  - the innovation's requirements for skills and expertise beyond those already within the organization;
  - the organizational restructuring needed to accommodate the new technology;
  - the availability of necessary skills and resources;
  - the organization's history and culture for change.
- Even with identical prior conditions and implementation objectives, implementation processes may vary considerably among adopting organizations (i.e., equi-finality). (Leonard-Barton, 1988).
- Conversely, the same innovation may be adopted by different organizations to fulfill widely divergent objectives.

These conditions limit the comparability of implementations, even where similar technologies are adopted. As a result, it is difficult to form a clear picture of the critical conditions that lead to successful technology implementation. Moreover, determining a project's "success" may pose serious problems. This is especially true for projects with multiple objectives. First, a new process technology's impact can be subtle and widespread, extending well beyond the altered operation. For example, the installation of a robotic system may reduce labour costs for the specific application as intended, but increase material costs (because of the need for more consistent inputs) and indirect costs (because higher skills in programming and maintenance may be required). Additionally, the adoption of advanced equipment may affect worker attitudes, impact sales or even accompany a complete shift in corporate capabilities. Second, few firms gather comprehensive information about completed projects, with post-audits done poorly, if at all. As a result, there are usually scant data about the new technology's actual performance against the criteria – such as improved efficiency or enhanced profitability – by which the project was initially justified.

Under these circumstances, investigators have to use the best information that is generally available to compare relative success among implementation projects. This means that some surrogates of inputs and initial conditions have to be used and that outcomes have to be judged on a broader basis, with multiple criteria reflecting the many objectives managers have for their particular projects. In examining the implementation of one type of technology – computerized information systems – Pinto and Slevin (1988) propose three criteria for assessing project outcomes: technical validity (whether the technology works as intended), organizational validity (whether clients or users are satisfied with the outcomes) and organizational effectiveness (whether the organization achieved overall positive benefits).

Most new technology projects can be assessed for their technical and organizational validity; however, measuring organizational effectiveness is much more difficult. It is difficult to attribute the organization-wide net benefits of a process technology improvement, including changes in profits or market share, image enhancement, etc., even within a firm that conducted intensive, multi-year post-audits. It is virtually impossible to do so within those firms that conducted little or no follow-up work (which are the majority). Without the ability to gauge a project's organizational effectiveness, the measurement of the relative success of a technology implementation is based almost exclusively on operational measure of technical validity and organizational validity.

Comparing outcomes is also complicated by the differences among projects and the initial conditions at each site. Any comparison of relative project success must control for those conditions that would be expected to impact eventual results. Two dominant conditions that often vary from site to site are the project's technical design and the initial availability of critical technical skills and resources. Most process technologies are at least in part unique, since they must be modified to suit the particular requirements of each plant's operations; in some cases, they may be virtually custom-made for each application. The levels of modification impose different challenges, with differing levels of system complexity and differing dependence on newly developed hardware and software. Sites also vary in their access to experienced personnel who would have the technical and management expertise. It is only appropriate that comparisons of outcomes should be made relative to these input conditions. The DEA methodology accomplishes the requirements of comparing projects with differing initial conditions and with various goals by using multiple constraints or input conditions to determine relative success (or efficiency) ratings of comparable projects on the basis of numerous outcome measures.

12.2.2 Assessing the Implementation of Industrial Robotic Systems

To demonstrate the comparison of similar implementations, we use a group of projects that installed robotic systems in industrial plants. In each case, the system was the plant's first use of robots. Information on each implementation is taken from project records, interviews with a number of managers at each site and direct observations of the systems in action.

Few of the 31 plants had any detailed records about actual costs for their projects, either for their system's initial installation or for their subsequent operation. Costs of purchased equipment and services were the only expenditures that all sites could report, even though all of the plants incurred other costs – for modifications, training, plant preparation and management time – that were usually substantial but rarely captured. In some cases, the non-capital startup costs were estimated to have been very high and continued for several months following system installation. Cost, then, provided little indication of the inputs actually required to complete these projects.

The project outcomes were equally difficult to judge on the traditional basis of efficiency improvement. For example, 14 of the 31 projects were welding applications; with these projects, the main benefits included not only reduced cost but improved weld quality plus the automation of hazardous operations (where, for instance, the increased automotive use of galvanized metal introduced new problems with dangerous fumes in welding). Moreover, in many different applications, robotic systems failed to reduce labour costs as expected, because the robots were found to require constant monitoring. Despite the lack of clear economic advantage in many projects, almost all of the systems became routinized. With some smaller firms, the robot systems were clearly inefficient compared to the direct costs of doing the same work manually, but the firm managers believed that the improved consistency and the enhanced image gained by adopting robots made up for the higher costs.

Given these conditions, measures that indicated the projects' technical and organizational validity were judged to be most appropriate for gauging outcomes, rather than comparing financial benefits. Objective measures of how well the system worked, both initially and in routine use, were used to indicate its technical validity. Subjective measures (managers' opinions) of how well the system achieved expectations indicated organizational validity. More specifically, technical validity was gauged by:

- *startup duration*, the length of time from system installation until the completion of debugging;
- *uptime*, the proportion of the scheduled running time that the system

was capable of operating once in routine use.

Organizational validity was judged by:

- *management satisfaction*, the perceived degree to which the project met expectations.

Management satisfaction reflected multiple opinions wherever possible, with the mean score being used as the measure.

Conditions under which the projects were carried out varied considerably from site to site. Some of the firms that adopted robots were small plants with limited technical resources and experience with programmable equipment; others were large plants, with highly competent technical staff and considerable experience using and maintaining other forms of programmable equipment. A measure of previous technical experience with similar technologies was developed to differentiate these conditions in each plant. Additionally, the systems themselves varied widely in characteristics that were likely to affect implementation ease. Some systems were relatively simple, consisting of a single robot interacting with one other piece of automated equipment; other systems contained up to twelve robots that interacted with several other machines. Some systems were designed to perform a single task while others had to be programmed to carry out numerous different operations, or perform similar operations on a wide variety of different work-piece designs. These design elements were scaled and combined to form a measure of each system's complexity.

The systems also varied in the degree to which they employed newly developed technology. Some systems employed proven, "off the shelf" robots, controllers and other components while others required specially - prepared grippers, part-handling mechanisms and other machines. Since newly developed manufacturing equipment tends to face more startup problems and failures, a measure of this facet of each system was devised to control for this condition.

Appendix 1 contains a brief description of each of the input and outcome measures. Some of these measures are interval; others are only ordinal. For some of the measures, the derived values form distinctly non-linear patterns, while others have limited variability. These conditions severely limit the usefulness of least-squares-based analytical techniques. Fortunately, the DEA methodology can deal with these problematic measurements.

12.2.3 The Model Structure

A robotics project can be viewed as a Decision Making Unit, in much the same manner as one would view a factory, government department, etc. It utilizes resources or inputs, and produces outputs or results. Comparison of a project to an operating unit is a useful paradigm, if one considers the manner in which the productivity standing of a factory would be measured. In such an industrial setting, efficiency or productivity is generally approached from an *engineering* perspective. This approach is based upon production *standards*. Such standards specify the best (minimal) amount of each input needed to produce one unit of output, and the productivity rating of the DMU is generally given by the ratio of standard or required inputs to actual inputs consumed.

While productivity standards *can* be derived in an engineering environment, such is not the case in the project setting. Outputs or outcomes are not reducible to a common unit of measurement such as hours or dollars, and inputs are not purely economic, but rather characterize the environment or circumstances surrounding the implementation. Thus, an alternative to the *absolute* approach to efficiency measurement offered by the conventional engineering method, is a *relative* efficiency model (DEA). For our purposes here we adopt the constant returns to scale version of the DEA model.

There are several important reasons why this method is particularly well suited to the implementation efficiency problem at hand. First, factors such as technical complexity of the new system and its use of new technology form an important part of the picture, yet are not easily reducible to economic units for purposes of setting standards. Therefore, as indicated, the conventional approaches are not applicable here. At the same time, scale measurements of these factors *are* available, allowing for a relative comparison of the projects. Second, the method possesses certain characteristics which render it a valuable tool in the context of the present problem. These characteristics are:

1. The model can provide a clear discrimination among the projects, hence separating them into various rank classes;
2. It can help in pointing to reasons for apparent inefficiencies, therefore aiding in either verifying or disproving popular belief vis- à-vis influences on implementation success;
3. The model makes allowance for any special circumstances prevailing in some project settings (e.g., low versus high degree of technical or management experience); and
4. Its structure is such that one can evaluate parameters that are not directly included in the model, yet which may have an important impact on implementation efficiency (e.g., number of employees in plant, urgency of the project, etc.).

This approach for determining weights for each DMU can be justified by arguing that since non-economic factors are present, there is no "correct set" of weights that apply to all projects. The importance, for example, of the complexity factor to projects in large plants may be different than to those in small plants. Thus, rather than having to try to assign some set of common weights to inputs and outputs, the model itself chooses weights that are appropriate for each project.

In the sections to follow, we elaborate on the problem setting and the outcomes resulting from the DEA analysis.

12.3. THE DATA

The DEA evaluation of implementation efficiency was applied to data collected on 31 robotics projects at 30 different sites implemented during the period 1980 to 1986, as detailed in McCutcheon (1988). In each case, the project was the plant's first use of robotic technology. In one case, two projects were installed simultaneously and independently. Data were collected by onsite interviews, conducted primarily with process or manufacturing engineering managers. The distribution of the plant sizes, industrial sectors and uses for the robots are shown in Appendix 2.

The projects involved wide ranges of plant sizes, robot applications, system complexity levels and outcomes. Although it was anticipated that only successful projects would be found, two of the 31 projects studied had been abandoned and several were considered by their implementers to be partial failures.

Although arc welding applications and vehicle component manufacturing appear to dominate the sample, these categories in fact included widely varied projects. The vehicle component applications ranged from light stampings for automobiles to heavy welded sections for off-road equipment and military vehicles. While 42% of the surveyed systems were used for arc welding, these applications ranged from simple systems used in job shops to extremely complex systems dedicated to mass production.

The site studies resulted in the collection of data on a wide range of factors. For purposes of measuring implementation efficiency, six direct factors (3 outcomes and 3 influence variables as described earlier) were selected for use in the DEA model. In addition, three control variables were chosen to be used for further analysis. Appendix 1 contains a detailed description of these 9 variables. A table of the numerical values for the 31 sites investigated is contained in Appendix 3.

12.4. ANALYSIS OF EFFICIENCY

12.4.1 Preparing the Model for DEA Analysis

The implementation efficiency model described in section 2 was applied to the 31 projects. The effect of DEA analysis on the model is illustrated in Figure 1. Part of the input to this model is a set of limits on each of the three input and three output variables. The choice of limits is rather arbitrary. For purposes here, the limits on all inputs and on the output variable MSAT were set very wide based on the belief that there may be a high degree of uncertainty or ambiguity regarding these variables.

STIME and UPTIME were regarded as the variables whose data was perhaps the most reliable, and whose importance weights should be subjected to the tightest control.

12.4.2 Outcome from the Overall Analysis

The DEA model was run for the 31 sites, and an efficiency rating was obtained for each. Table 12-1 displays the results. In this case 7 of the 31 projects obtained an efficiency score of 1.00 or 100%, meaning that they are not dominated by other more efficient sites. The remaining 24 projects obtained ratings at a lower level, indicating that each is dominated by some other efficient site or combination of sites. As an example, compare sites #5 and #30. Site 5 had outputs that were higher than those for #30 (#5 took less time to put the project in place, its system is serviceable more often, and management satisfaction is higher than is true for #30). On the other side, the influences or inputs for #30 are more favorable than for #5 (meaning that #30 should have actually had *better* outputs than those of #5).

The outcomes in Table 12-1 demonstrate a wide range of ratings, meaning that the DEA model has been able to discriminate clearly among the implementation sites. Moreover, through arguments similar to that of the previous paragraph, one can detect, in many cases, clear reasons why some sites score low on the efficiency scale.

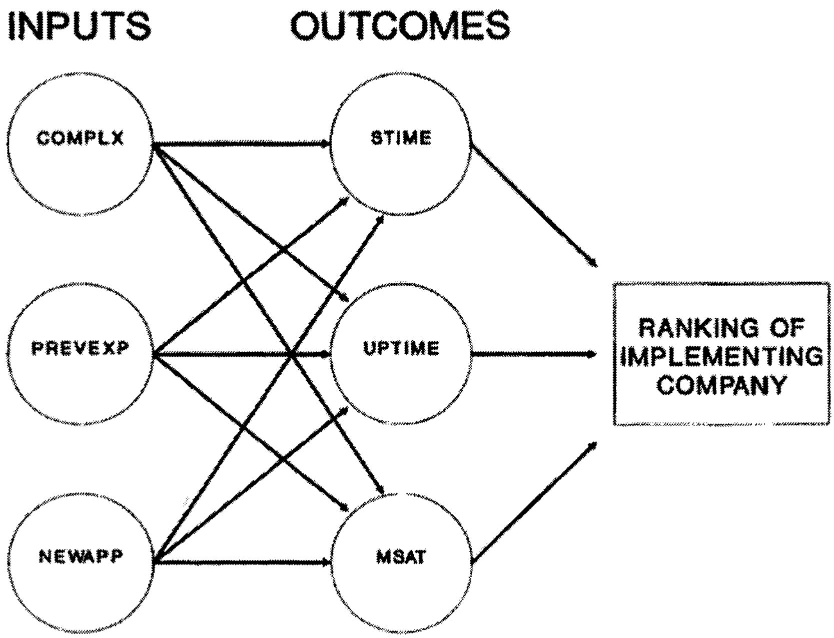

Figure 12-1. Implementation Efficiency Model

Table 12-1. DEA Efficiency Ratings for 31 Sites

Robotics Project	DEA Rating	Robotics Project	DEA Rating
1	62.9%	16	66.6%
2	81.1%	17	69.0%
3	70.9%	18	100.0%
4	52.4%	19	52.2%
5	100.0%	20	64.4%
6	64.1%	21	100.0%
7	52.3%	22	100.0%
8	82.2%	23	69.3%
9	69.0%	24	93.0%
10	54.9%	25	100.0%
11	69.5%	26	58.6%
12	100.0%	27	87.8%
13	65.0%	28	100.0%
14	65.2%	29	57.2%
15	51.6%	30	58.7%
		31	89.4%

12.4.3 Using the Control Parameters

Three additional parameters — plant size, the use of supplier management techniques, and the project's perceived urgency – were examined for their impacts on implementation efficiency. Each parameter was expected to influence results. Comparing the initial efficiency ratings against those derived by including each of these parameters shows how rankings would be affected if the parameter were to be considered an input.

Size has been found repeatedly to correlate with organizational innovativeness, with larger organizations tending to adopt more innovations. (Ettlie and Rubenstein, 1987). However, as pointed out by Rogers, size is most likely a surrogate for other factors, such as organization structure or the availability of slack resources, that have a more direct influence. (Rogers. 1983). During the implementation phase, plant size may have two effects: larger plants may have more sophisticated staff (reflected in the measure for previous experience with similar technologies) and more sophisticated infrastructures, which should lead to better results; (Gerwin. 1988), at the same time, larger plants may have more mechanistic managerial relationships which may *impede* implementation efficiency.

The use of a wide range of supplier management techniques for controlling technical innovation projects conducted largely by vendors has been shown to reduce system startup problems significantly. (Wood and Elgie, 1976). Including a measure of each plant's supplier management methods in the analysis provides an indication of their impact on overall implementation results.

The urgency associated with getting these robotic systems into routine use varied considerably from project to project. In some cases, the new systems were seen as essential for assuring the plant's continued survival, whereas in others, the systems had no special urgency, being viewed in part as experimental, undertaken as learning opportunities. Perceived urgency has been shown to have an influence on the initial stages of innovation adoption decisions. (Ettlie and Vallenga, 1979). Urgency was included here to see if it had a discernible impact on the project's implementation stage as well.

To gain some insight into the influence of a parameter such as supplier management, the 31 project sites were separated into two groups. Those sites with a supplier management rating between 0 and 8 (*low*) were separated from those with a rating higher than 8 (*high*). A DEA analysis was then applied to each of these groups separately. Table 12-2 shows the outcomes. Column #1 displays the original ratings prior to the split off analysis (same ratings as in Table 12-1). Column #2 presents the *new* ratings for those projects in the low supplier management group (15 projects), while column

#3 shows the revised ratings for the high supplier management group (16 projects).

Table 12-2. Evaluation of Supplier Management

Robotics Project	DEA Rating	Low Sup Mgt. Group	High Sup Mgt. Group
1	62.9%	72.7%	
2	81.1%		81.1%
3	70.9%	71.7%	
4	52.4%	71.7%	
5	100.0%		100.0%
6	64.1%		73.1%
7	52.3%	64.1%	
8	82.2%		90.6%
9	69.0%		77.4%
10	54.9%		58.5%
11	69.5%		70.9%
12	100.0%		100.0%
13	65.0%		70.8%
14	65.2%		70.7%
15	51.6%	59.6%	
16	66.6%	81.4%	
17	69.0%	87.4%	
18	100.0%		100.0%
19	52.2%	59.1%	
20	64.4%		75.3%
21	100.0%		100.0%
22	100.0%		100.0%
23	69.3%	69.5%	
24	93.0%	93.1%	
25	100.0%	100.0%	
26	58.6%		71.4%
27	87.8%	100.0%	
28	100.0%	100.0%	
29	57.2%	85.1%	
30	58.7%		62.5%
31	89.4%	89.8%	
	Mean Score	80.3%	81.4%

It is noted that the rating for a project in any subgroup analysis is always at least as high as is the value for that project when the entire group is under consideration. For example, the rating for project #1 increased from 62.9% to 72.7% when the low supplier management group was split off for a separate analysis. The reason for this is that the project (#1) is being compared to a smaller group in the latter case.

Table 12-3 presents summary statistics, specifically the arithmetic means, for the two subgroups; *before* and *after* the split.

When members of the low supplier management group were being evaluated relative to all 31 projects, the average rating for this subgroup of 15 sites was 71.6%. When this subgroup of 15 was split off for separate analysis, the average rating grew to 80.3%. The corresponding figures for the high group were 77.0% and 81.4%.

Table 12-3. Average Efficiency Ratings of Low and High Supplier Management Groups

	Supplier Management	
Analysis Group	Low Scoring	High Scoring
Before	71.6%	77.0%
After	80.3%	81.4%
Change in Average Score	8.7%	4.4%

At least three important observations can be made regarding the high versus the low supplier management group. First, it is observed that the low group has a worse performance standing on average than is true of the high group (71.6% versus 77.0%). The second observation is that the low group's average increases by 8.7% (80.3-71.6), while the high group average increases only by 4.4%. It can be argued that the original efficiency gap of 28.4% (100% - 71. 6%) for this low scoring group is closed by 8.7% due to the removal of high supplier management sites. Thus the size of the gap removed representing the change in the average score in Table 12-3 is a reflection of how one group suppressed the scores of the other. Since the reduction in the high scoring group efficiency gap (4.4%) is less than that of the low scoring group (8.7%), the high scoring group had more of an effect on keeping the low scoring group's rankings low than vice versa. The third observation is that there were more top performers in the high group than in the low, prior to the split off. Specifically, in the set of *seven* sites which achieved an efficiency score of 100%, *five* of these were in the *high* supplier management group, while *two* only were in the *low* supplier management group.

Hence, it can be said that for the particular sites in question, a *high* degree of supplier management appears to have a *positive* influence on implementation efficiency.

Tables 12-4 and 12-5 present summary statistics similar to those of Table 12-3 for plant size and urgency, showing the arithmetic means before and after being split into two subgroups.

Here again, the group with the smaller sites has the higher average rating (when compared to the larger sites), contains the majority of the top performers, and views the opposing (larger sites) group as presenting little or no effect on their scoring. For the large plants, quite the opposite is true. The average rating is very low (69.7%), two top performers only are present, and an enormous effect is removed when the group of smaller plants is eliminated from the comparison set. That is, given that the efficiency gap of 30.3% (100%-69.7%) is closed by 15.9% (85.6%-69.7%), one can argue that this gap is *explained* by the opposing group's presence (or lack of presence).

Table 12-4. Average Efficiency Ratings of Small and Large Sites

	Plant Size	
Analysis Group	Small (0-200 Employees)	Large (Above 200 Employees)
Before	79.5%	69.7%
After	80.0%	85.6%
Change in Average Score	.5%	15.9%

Table 12-5. Average Efficiency Ratings of Low and High Urgency Sites

	Urgency	
Analysis Group	Low Scoring	High Scoring
Before	73.3%	77.2%
After	73.3%	91.5%
Change in Average Score	0%	14.3%

The outcomes for the urgency parameter are somewhat difficult to interpret. First, it should be noted that the scale values for this parameter ranged for 1-4, which may not provide a measure with sufficient variability. Furthermore, an attempt to split the 31 sites into two relatively equal sized groups failed. There were 22 low urgency (i.e., those ≤ 2 on the 1-5 scale), and only 9 high urgency (i.e., > 2 on the 1-5 scale) sites.

The fact that the average rating in the low urgency class did not change from before to after the split off is due to the fact that, while not all of the top performers lie in the low urgency class, *those* top performers against which the *low* urgency sites are measured are only those in the low group. That is, none of the low urgency sites are being compared to the two top performers that fall in the high urgency group.

On the other hand, since the average rating for the high urgency group does change (moves from 77.2% to 91.5%), clearly a number of the members of this group were being evaluated against top performing low urgency sites. Note, for example, that prior to the split off, the high urgency site #2 had a rating of 81%. When the split off occurred (the low urgency group was removed), its rating climbed to 100%.

While most of the top performers fall in the low urgency class, it is true that the worse performers also occupy this class (e.g., sites #4, 7, 15 and 19). Thus, the low urgency group is very heterogeneous. The high urgency group is less so. Little else can be concluded about the urgency impact on implementation efficiency.

12.5. DISCUSSION AND SUMMARY

Each of the plants had its own set of criteria for judging the success of its project. Some projects were designed specifically to reduce labour costs; others were economical means to reduce workplace hazards; still others were viewed, at least in part, as experiments in flexible technology applications. On these bases, the projects had limited comparability. However, at a fundamental level, all projects were expected to meet common criteria of technical success and organizational acceptance.

Using measures for these common criteria, the DEA methodology provides a means of ranking the projects according to their outcomes. The analysis using DEA has two advantages. First, it indicates the relative achievements of the plants, given the widely differing conditions they faced in implementing their first systems involving robots. Moreover, it allows this comparison in an environment where the data made parametric analytic tools inappropriate. DEA has a major advantage in allowing the use of data as they are found in the real world.

Second, the model was capable of showing the impact of particular conditions on relative implementation efficiency. One of these conditions, management's perception of the project's urgency, had results which proved difficult to interpret. However, the other two, plant size and the use of supplier management techniques, proved to be conditions that had to be taken into account when assessing the relative efficiencies of the plants. When comparing the implementation efficiency of plants, managers must judge them relative to plants of similar size and supplier management capability. Interestingly, in the short term, good supplier management is within the project manager's ability to influence, while plant size is a factor usually determined by the firm's top management. Therefore, managers assessing relative project performance may choose to disregard the former factor while considering the latter as a basis of comparison.

Advanced manufacturing projects such as the introduction of robotic systems into individual plants will continue to be evaluated using traditional financial measures such as return on investment. Unfortunately, as often noted, these metrics will not always capture the full impact of new technology on a plant's competitiveness. Inevitable errors in the forecasting of cash flows and parameters such as discount rates will produce comparisons of planned to actual performance that are difficult to interpret. As noted by Kaplan and Johnson (1987) traditional cost accounting systems are struggling to furnish better information needed to evaluate projects. There will always be a need to have a common financial measure such as ROI to compare proposed projects competing for scarce resources. This should not be confused with the need for effective post-project assessments of project performance that contribute to organizations learning about how to implement better.

We would recommend that project gatekeepers or evaluators expand the number and variety of types of measures for evaluating advanced manufacturing project performance. Tools, such as DEA, facilitate the resolution of the resulting "messy" evaluation task. The ability to parse the effect of controllable factors such as supplier relationships versus less controllable factors such as plant size, leads to fairer assessments of individual project leaders and their team members. Consistent project evaluation techniques would provide metrics for a corporate wide database of innovation aimed at speeding transfer of best practices between plants in large companies. (Johnston and Leenders, 1990). Post project audits can be made more precise in their identification as to which operating and organizational policies should be changed to facilitate project implementation. In addition, private and public agencies responsible for evaluating research output from diverse projects can more consistently rank projects under their control.

Given the problems in comparing various new technology implementations, the DEA technique provides an effective means of judging project outcomes. Its ability to use widely differing forms of data as indicators in inputs and outputs helps to overcome the inherent difficulties with such comparisons. In addition, using a database such as the one here, managers can determine the appropriateness of considering certain factors in judging project results. In this way, projects can be compared relative to others that faced similar conditions, providing potentially fairer comparisons.

REFERENCES

1. Cook W., Y. Roll and A. Kazakov (1990), A DEA model for measuring the relative efficiency of highway maintenance patrols, *INFOR* 28, 113-124.
2. Ettlie J. and A. Rubenstein (1987), Firm size and product innovation, *Journal of Product Innovation Management* 4, 89-108.
3. Ettlie J. and D.B. Vallenga (1979), The adoption time period for some transportation innovations, *Management Science* 25, 429-443.
4. Gerwin D. (1988), A theory of innovation processes for computer-aided manufacturing technology, *IEEE Transactions on Engineering Management* 35, 90-100.

5. Johnston D.A. and M.R. Leenders (1990), The diffusion of innovation within multi-unit firms. *International Journal of Operation and Production Management* 5, 15-24.
6. Kaplan R.S. and H.T. Johnson (1987), Relevance lost: the rise and fall of management accounting, *Harvard Business School Press*, Boston.
7. Leonard-Barton D. (1988), Implementation as mutual adaptation of technology and organization, *Research Policy* 17(5), 251-267.
8. McCutcheon D.M. (1988), The role of technical skills in the implementation of robots, Doctoral Thesis, The University of Western Ontario, London, Canada.
9. Pinto J.K. and D.P. Slevin (1988), Critical success factors across the project life cycle, *Project Management Journal*, 19(1), 121-135.
10. Rogers, E., (1983) The Diffusion of Innovations (3rd Ed.). The Free Press, New York.
11. Wood, A.R. and R.J. Elgie (1976), Early adoption of manufacturing innovation, Research Report, The University of Western Ontario, London, Canada.

This chapter is based upon Cook, W.D., D. Johnston, and D. McCutcheon, (1992), Implementation of Robotics: Identifying Efficient Implementors, OMEGA – International Journal in Management Science, 36(3), 177-188, with permission from University of Toronto Press.

APPENDIX 1: VARIABLES IN IMPLEMENTATION EFFICIENCY MODEL

Three initial conditions at the start of a project:

SYSTEM COMPLEXITY (COMPLX): A count of four components of robotic systems, summed to created an absolute scale of overall system complexity. The four components are:

- the number of machines controlled by the central controller
- the number of unique part numbers that require distinct programming
- the number of robots
- the number of discrete operations performed by robots in a cycle (e.g., lift, load, perform one weld)

PREVIOUS EXPERIENCE WITH TECHNOLOGY (PREVEXP): This variable captures the expertise of the production department relevant to the new system. It is a summed score of four 5-point measures that were scaled for;

- number of system-years operating programmable equipment
- maintenance capabilities with programmable equipment
- number of system-years of experience with systems powered by the same method
- maintenance expertise with similar mechanical systems.

NOVELTY OF THE APPLICATION (NEWAPP): was based on 5-point scales for each major component; each scale reflecting the component's innovativeness ranging from the purchase of a standardized off the shelf equipment to the development of a customized component.

- Three outcomes at the end of the project:

START UP TIME (STIME): the number of weeks required to take the technology from physical installation to the point where it was routinized into normal production. Routinization was determined by either the point of official hand over of the technology to day to day operations or the date of the last major modification.

UPTIME (UPTIME): an estimate of the percentage of the total production time available for operation in which the technology was in service.

MANAGEMENT SATISFACTION (MSAT): a perceptual measure on a 5 point scale from very dissatisfied to highly satisfied with the degree to which the project met expectations.

- Control Parameters:

SUPPLIER MANAGEMENT (SUPMGT): This is a measure which rates the project teams efforts to manage their system suppliers. It has 7 categories:

- use of written specifications for the work to be done
- requirements that the vendor provide adequate documentation prior to the system's final installation
- continual contact during the equipment development stages (e.g., weekly or bi-weekly meetings or telephone calls)
- visit(s) to vendor's plants while equipment was being built
- requirements for pre-tests under plant conditions
- prior assessment of vendor's financial strength
- prior assessment of vendor's technical capability

Each category was scored 0 (not carried out), 1 (carried out, but only for this special project) or 2 (carried out; routine procedure for the plant or for the project manager).

PLANT SIZE (PLTEMP): this variable records the number of employees within the plant when the system was implemented.

URGENCY: (URGENCY) this measure is on a 5-point scale, based on the reporting manager's assessment of the urgency connected with the project.

APPENDIX 2: SITE DEMOGRAPHICS

Plant Size

Number of Employees at the time of the project	Number of sites
up to 100	5
101 - 250	11
251 - 500	8
more than 500	7

Robot System Applications

Primary use of the robot(s)	Number of projects
Arc welding	13
Part handling	6
Machine loading/unloading	5
Assembly	2
Spot welding	1
Soldering	1
Palletizing	1
Toolhandling	1
Adhesive application	1

Industry Sector

Industrial sector served by the project	Number of projects
Vehicle assembly	2
Vehicle component manufacture	11
Heavy engineered equipment	1
Metalworking job shop	2
Electronics component assembly	2
Appliance assembly	1
Construction materials	1
Plumbing fixtures and supplies	3
Other commercial/industrial products	6
Pharmaceutical laboratory work	1

APPENDIX 3: DATA MATRIX OF VARIABLES AND PROJECTS

Robotics							Control		
Project	Inputs				Outputs		Parameters		
	NEWAPP	PREVEXP	COM-PLX	MSAT	UP-TIME	S-TIME	URGENCY	PLTEMP	SUP-MGT
1	11	11	39	4	78	197	2	160	4
2	4	20	27	2	95	184	3	3500	10
3	14	6	39	4	85	175	2	76	7
4	13	18	33	4	78	150	1	300	7
5	11	4	38	5	90	188	3	350	10
6	11	9	40	4	78	176	1	245	13
7	16	16	42	3	85	194	1	300	5
8	10	9	36	3	97	180	2	160	13
9	8	12	34	4	80	188	4	1200	12
10	9	19	34	3.5	78	176	2.5	500	11
11	5	18	39	5	100	176	2	300	10
12	4	15	14	4	78	144	3	200	9
13	11	18	34	5	95	176	3	1700	11
14	13	16	38	4	99	199	1	40	12
15	16	13	42	3	78	188	1	120	7
16	15	11	28	1	65	184	3	1100	8
17	15	11	28	3	78	188	3	1100	8
18	2	11	31	4	90	188	1	120	18
19	17	14	42	4	78	196	2	200	8
20	9	12	36	2	78	190	1	520	10
21	2	11	22	4	78	185	2	120	13
22	8	16	2	2	78	135	1	600	10
23	19	4	34	3	78	100	1	200	6
24	11	4	38	3	80	195	2	75	4
25	5	6	25	4	65	152	1	200	8
26	15	14	35	2	83	189	2	500	11
27	3	11	34	3	80	179	1	150	5
28	10	4	25	5	78	198	2	60	8
29	5	18	36	4	78	145	2	80	5
30	10	16	32	2	78	165	1	450	10
31	9	4	28	4	40	174	3	280	7

Chapter 13

SETTING PERFORMANCE TARGETS FOR NEW DMUS

13.1. INTRODUCTION

A problem of considerable interest to many organizations, and to be examined herein, involves the setting of *performance* targets for a *yet to be created* decision making unit. Typically, such a problem arises in site selection decisions for new *facilities*. Consider, for example, the selection of a site for a health care facility (clinic, hospital, etc.). Suppose that estimates for the demand for various types of services - for example, geriatric care, prenatal services, emergency provisions, and so on — have already been established. That is, the outputs y_{rj} are given values, or at least can be estimated. An issue affecting the design of the facility is that involving the inputs. While some inputs are given values, such as the demographics of the population where the facility is to be located, other factors may be less certain. More to the point, the resources, such as staffing needs, required in order to be able to deliver those services are at the discretion of the organization. One approach to addressing the resources side of the problem is to set as targets those staffing levels which would ultimately result in an efficiency score for the new facility that is at or above some acceptable level. Stated in DEA terms, the *resource targeting* problem is one in which the outputs (services) are given or known, while some of the *target* inputs (staffing, operating budgets, etc.) are values to be chosen in such a way that the relative efficiency rating meets some desired standard. Hence some inputs are discretionary while others may be nondiscretionary.

In other situations, the roles of outputs and inputs may be reversed. In the site selection exercise for a retail establishment, the costs or input parameters may be known, or can be estimated – size of facility, staff makeup, operating budget, etc. The problem in this case becomes one of setting *output targets*, such as sales of various products or services, that will result in an overall performance measure, i.e., efficiency score, that meets some acceptable standard. An example might be the positioning of a new branch of a bank. If in a given setting, the approximate cost or input values are known, the bank may want to establish sales targets for the branch for various loans, investment sales in GICs and RRSPs, etc. These targets, which would result in an efficiency rating at or above some level, can then be compared to estimated outputs to see if the proposed site is desirable and can meet expectations. Viewed from this perspective, the output target problem is pertinent to the marketing of new products and/or the establishment of new sales territories.

13.2. THE PERFORMANCE TARGET PROBLEM WITH NO RESOURCE BOUNDS

13.2.1 The Input Target Problem

For a given set of n decision making units, for example, health care facilities, let $\{y_{rj}\}_{r=1}^{R}$ be the set of given outputs and $\{x_{ij}\}_{i=1}^{I}$, the set of inputs for facility j. The *input oriented* DEA model presented earlier can be utilized to derive an efficiency score for each DMU j. Specifically, we adopt the BCC model (L-P version) for purposes of discussion here: Model (13.1) is the dual and (13.2) the primal form of the model.

$$\begin{aligned}
&\max_{\mu,\nu,w} \sum_{r=1}^{R} \mu_r y_{rjo} + w \\
&\text{subject to} \\
&\sum_{i=1}^{I} \nu_i x_{ijo} = 1 \\
&w + \sum_{r=1}^{R} \mu_r y_{rj} - \sum_{i=1}^{I} \nu_i x_{ij} \le 0, j = 1,\dots,n \\
&\mu_r, \nu_i \ge 0, \text{all } r, i,\ w \text{ unrestricted.}
\end{aligned} \tag{13.1}$$

and

$$\min \theta$$
subject to

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{rjo}, r = 1 = 1, \ldots, R \tag{13.2}$$

$$\theta x_{ijo} - \sum_{j=1}^{n} \lambda_j x_{ij} \geq 0, i = 1, \ldots, I$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j \geq 0, \ \theta \text{ unrestricted}$$

Consider now the problem of selecting a site for a new $(n+1)$st facility, the performance of which must ultimately be judged against the other existing DMUs. For present purposes, we assume that the demand for services, for example, the outputs $\{y_{rn+1}\}_{r=1}^{R}$ are given values. Suppose, however, that the inputs are not immediately available, and that the problem at hand is to set *target* values $\{x_{in+1}\}_{i=1}^{I}$ for these inputs, such that the resulting efficiency rating θ for this new facility is at least some acceptable level $\bar{\theta}$. To an extent, this problem addresses a feasibility issue. Specifically, the setting of an *efficiency performance floor* value of $\bar{\theta}$ permits one to determine the maximum inputs x_{in+1} (for example, operating expenses, staff levels, capital costs, etc.) that are allowable if the facility is to meet that value. If the actual estimated input requirements exceed the maximum limits, then presumably the venture would not meet the feasibility specifications.

Viewing the input target problem from the perspective of problem (13.2), the requirement is to find an input vector ($x_{1n+1}, \ldots, x_{In+1}$) so that the solution to (13.2) with $j_o = n+1$ gives an optimal θ value in excess of $\bar{\theta}$, that is we want $\hat{\theta}(= \min \theta) \geq \bar{\theta}$. In addition to the constraints given, additional restrictions may be imposed on the $\{x_{in+1}\}_{i=1}^{I}$. For example, there may be minimum or maximum levels x_{in+1} which some of the components x_{in+1} must obey. This case is addressed in Section 13.3.

Clearly, if the x_{in+1} represent resource requirements, such as staffing, then the larger their values, the more likely it will be that the outputs y_{rj} can be delivered. Thus, the larger the inputs the more flexibility the facility will have in conducting its operation.

To get a clear picture of the input target problem, consider a simple example of 2 DMUs, each producing a single unit of output and consuming 2 inputs in amounts:

DMU	Input 1	Input 2
1	10	2
2	12	1

We wish to find appropriate values x_{13}, x_{23} for a new DMU #3. Assume that an efficiency score of at least $\overline{\theta} = 80\%$ is to be achieved by the new DMU. Figure 13.1 illustrates the situation.

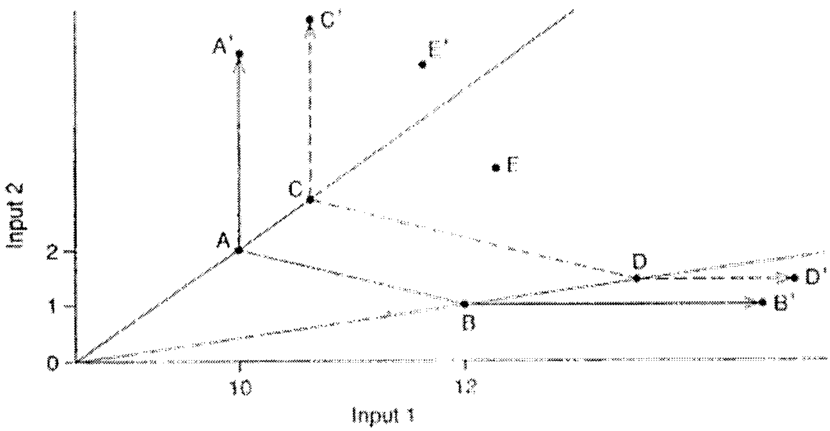

Figure 13-1. Isoquant for 2-D Problem

Points A and B show the positioning of the inputs for the two DMUs; the line segment AB is the only legitimate facet, and represents the efficient frontier. The portion of the cone enclosed by the two rays out of the origin and projected through A and B and lying behind the frontier represents that area where *properly enveloped* DMUs can lie (for example, at point E). A point such as E', outside the cone would be *improperly enveloped.* The line segment CD represents all possible properly enveloped DMUs that would have an efficiency score of $\overline{\theta} = 80\%$. That is, OA/OC = OB/OD = $\overline{\theta}$ = 0.80.

It is noted that as with the line segment AB, any points on the vertical line out of A (for example, point A') and on the horizontal line out of B (for example, point B') will have efficiency scores θ of 100%. Any improperly enveloped DMUs such as E' will be measured against these two frontier extensions, and will show positive slack in one of the second set of constraints in (13.2). Hence, any input target point (x_{13}, x_{23}) that has positive slacks will be improperly enveloped (for example, E'). All slacks must be zero if the point is to be properly enveloped.

In this simple case it is clear that any point inside the area ACDB is a candidate for the input target point (x_{13}, x_{23}).

The problem is then to derive a set of target inputs $\{x_{in+1}\}_{i=1}^{I}$ such that the resulting DMU will have an efficiency score $\theta \geq \bar{\theta}$, and will be as close as possible to being properly enveloped. We consider first the case where no additional restrictions such as upper and lower bounds are imposed on the $\{x_{in+1}\}_{i=1}^{I}$. This means that inputs can be reduced or increased as much as desirable without hitting any bounds.

13.2.2 Unimpeded Movement of Inputs

The problem to be solved is one of finding a set of target inputs $\{x_{in+1}\}_{i=1}^{I}$ that is maximal in some sense, such that the efficiency score $\theta^* \geq \bar{\theta}$ at the optimum of (13.2). The difficulty with dealing directly with this problem is that one has an optimization problem (min θ) within a larger optimization problem (max some function of the x_{in+1}). An alternative to this approach is to reverse the problem, i.e., find a *minimal* set of x_{in+1} while ensuring that $\theta \leq \bar{\theta}$. To this end, consider the following nonlinear problem:

$$
\begin{aligned}
&\min \alpha + \varepsilon \sum_{i=1}^{I} s_i \\
&\text{subject to} \\
&\alpha - x_{in+1} \geq 0, i = 1, \ldots, I \\
&\theta x_{in+1} - \sum_{j=1}^{n} \lambda_j x_{ij} - s_i = 0, i = 1, \ldots, I \\
&\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{rn+1}, r = 1, \ldots, R \\
&\theta \leq \bar{\theta} \\
&\sum_{j=1}^{n} \lambda_j = 1 \\
&\lambda_j \geq 0, \theta \geq 0, x_{in+1} \geq 0, \forall i, j.
\end{aligned}
\tag{13.3}
$$

The nonlinearity arises from the product of θ and x_{in+1}.

The term $\varepsilon \sum s_i$ in (13.3) guarantees that out of all alternate optima, any properly enveloped point ($x_{in+1}, \ldots, x_{In+1}$) will be chosen over an improperly enveloped point. In terms of the simple example above, the point will be chosen from within the area ACDB.

Lemma 13.1.
At any optimum ($\alpha^*, \lambda^*, x_{n+1}^*, \theta^*$) of (13.3), $\theta^* = \bar{\theta}$.

Proof:

If there exists an optimal solution to (13.3) for which $\theta^* < \bar{\theta}$ then there is a scaler $\gamma > 0$ such that the x_{in+1} values $\hat{x}_{in+1} = x^*_{in+1} - \underline{\gamma}$ are feasible in the second set of constraints in (13.3), when $\theta = \bar{\theta}$. Since all $\hat{x}_{in+1} < x^*_{in+1}$, then $\hat{\alpha} = \alpha^* - \gamma$ is feasible in the first set of constraints, meaning that α^* cannot be optimal. This contradicts the assumption on α^*, and the result follows.

Q.E.D.

From this lemma it follows that constraint $\theta \le \bar{\theta}$ may be removed and θ replaced by $\bar{\theta}$ in the second set of constraints, thus reducing the problem to a linear format. The next theorem follows immediately, and the proof is omitted.

Theorem 13.1:

If $(x^*_{1n+1}, \ldots, x^*_{In+1})$ is an optimal target input vector in the sense of problem (13.3), and if x_{ijo} and y_{rjo} are set equal to x^*_{in+1} and y^*_{rn+1} respectively in problem (13.2), then the optimum $\hat{\theta}$ in problem (13.2) is given by $\tilde{\theta} = \bar{\theta}$.

It is noted that at the optimum of (13.3), all $s_i = 0$.

From a technical efficiency standpoint, every point on CD is equally desirable in that they all produce a score of $\theta = \bar{\theta}$. Problem (13.3) is merely a vehicle for generating a feasible solution. From an economic or effectiveness standpoint, one would arguably wish to choose a least cost combination on CD if *prices* were known. Specifically, if $\{c_i\}_{i=1}^I$ were the costs of the inputs, it would be appropriate to choose $\{x_{in+1}\}_{i=1}^I$ according to the optimization problem

$$\min \sum_{i=1}^{I} c_i x_{in+1}$$

subject to

$$\bar{\theta} x_{in+1} - \sum_{j=1}^{n} \lambda_j x_{ij} = 0, i = l, \ldots, 1$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \ge y_{rn+1}, r = 1, \ldots, R \tag{13.4}$$

$$\sum_{j=l}^{n} \lambda_j = 1$$

$$\lambda_j, x_{in+1} \ge 0, \forall i, j.$$

The first set of constraints in (13.3) guarantee that only properly enveloped points are considered.

For the simple example presented earlier, point C will be the optimum to problem (13.3), whereas the optimum for (13.4) would depend upon the prices c_i.

13.3. RESTRICTED INPUT SPACE

If additional restrictions Ψ are imposed on the x_{in+1} (for example, Ψ is not the full space R^I), the nature of the feasible space as per Figure 13.1 can change significantly. We consider two types of restrictions – upper bounds and lower bounds on the x_{in+1}.

13.3.1 Upper Bounds on the x_{in+1}

In the case that upper bounds x^u_{in+1} are imposed on the allowable input targets, then Ψ is defined as

$$\Psi = \{(x_{in+1},\ldots,x_{In+1})\varepsilon R^I \mid x_{in+1} \le x^u_{in+1}\}. \qquad (13.5)$$

This definition allows, of course, for the case that some x^u_{in+1} may be ∞. A finite upper bound $x^u_{i_on+1}$ may, for example, exist for some input i_o if the organization simply imposes a limit on a particular resource (for example, in the case of a bank, the number of back office staff may not be permitted to exceed some level within any branch of the type being considered.) Alternatively, the maximum rent that might be paid in a particular location would in general be bounded.

If the additional constraints (13.5) are appended to (13.3), and if $\theta = \overline{\theta}$ is feasible, then an appropriate set of target inputs can be derived. If $\overline{\theta}$ is not feasible then no $\theta < \overline{\theta}$ will be feasible either. In this case, the lowest efficiency score $\hat{\theta}$ allowable for the $n+1$ st DMU will be some value strictly larger than $\overline{\theta}$. To determine $\hat{\theta}$, set all restricted x_{in+1} at their upper bounds x^u_{in+1} and let all unrestricted variables be assigned a large value M. Call these values $\overline{x}_{in+1}$. Then, solve the minimization problem

$$\min \theta + \varepsilon \sum_{i=1}^{I} s_i$$

subject to

$$\theta \overline{x}_{in+1} - \sum_{j=1}^{n} \lambda_j x_{ij} - s_i = 0, i = 1,\ldots,I$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{rn+1}, r = 1, \ldots, R \tag{13.6}$$
$$\sum_{j=1}^{n} \lambda_j = 1$$
$$0 \leq \theta \leq 1$$
$$\lambda_j \geq 0, j = 1, \ldots, n$$

To visualize the problem, a revised version of Figure 13-1 is shown as Figure 13-2, where an upper bound on $\underline{x}_{in+1}$ has been imposed. Depending upon the positioning of the bound, $\theta = \overline{\theta}$ may or may not be feasible. Since the bounding plane shown as the vertical line out of $\overline{x}_{in+1}$ does not interact the $\overline{\theta}$ bounding CD, then the optimal solution $\hat{\theta}$ will exceed $\overline{\theta}$. In this illustrative case, the optimum will occur at point F and $\hat{\theta}$ = OA/OF.

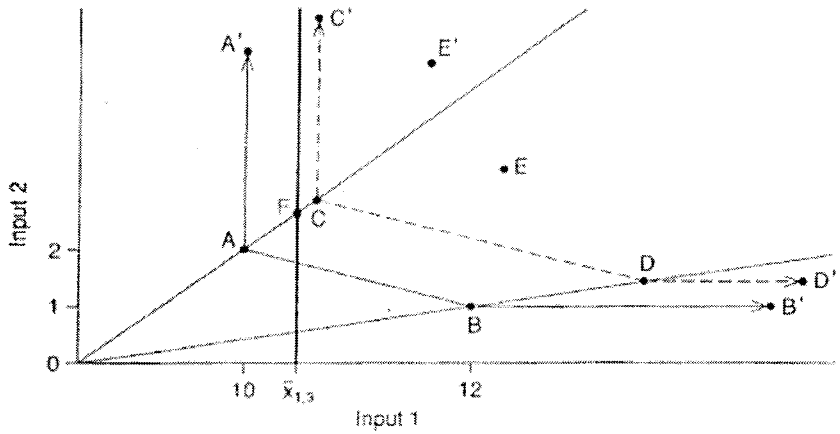

Figure 13-2. Imposed Upper Bound on Input 1

In general, with upper bounds, the optimum set of target inputs is given by

$$\hat{x}_{in+1} = \overline{x}_{in+1} - s_i^*, \tag{13.7}$$

where $\overline{x}_{in+1}$ is either an imposed upper bound x_{in+1}^u or equals M. s_i^* is the slack in the i th constraint of the first set of restrictions in (13.6).

Clearly, problem (13.6) will be feasible provided at least one point on the efficient frontier satisfies the constraints. In the example of Figure 13-2 this means that the upper bounding plane cannot be to the left of the segment AA . In this case θ would exceed 1.

13.3.2 Upper Bounds on the x_{in+1}

In the case that lower bounds x^L_{in+1} are imposed on some of the input variables, Ψ is defined by

$$\Psi = \{x_{in+1}, \dots, x_{In+1})\varepsilon R^I \mid x_{in+1} \geq x^L_{in+1}\}. \tag{13.8}$$

Again, it is noted that some lower bounds may be zero. The feasible region might now appear as in Figure 13-3, Figure 13-4 or Figure 13-5; these figures represent 3 possible situations regarding the degree of restrictiveness of the lower bounds. In Figure 13-3, where we require $x_{13} \geq \bar{x}_{13}$, a portion GB of the efficient frontier AB is in Ψ. This means that if we solve problem (13.3) with (13.8) imposed, then a set of target inputs will be found for which the projected point will be on GB. In this particular example the target input point will be at H. A DMU at H can be projected unimpeded directly to G. Any point between I and H would first project to the IG segment where input 1 would become nondiscretionary, and the projection would then be vertical (with only input 2 as discretionary now) to point G. This would result in a θ smaller than $\bar{\theta}$.

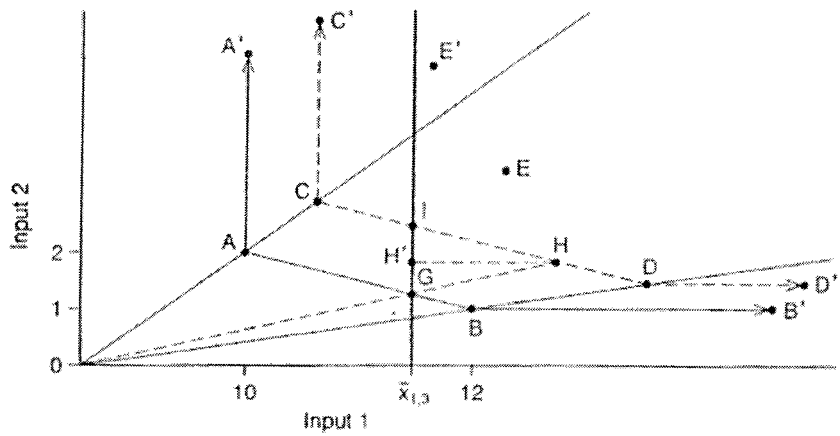

Figure 13-3. Lower Bound where $\phi^* = 1$

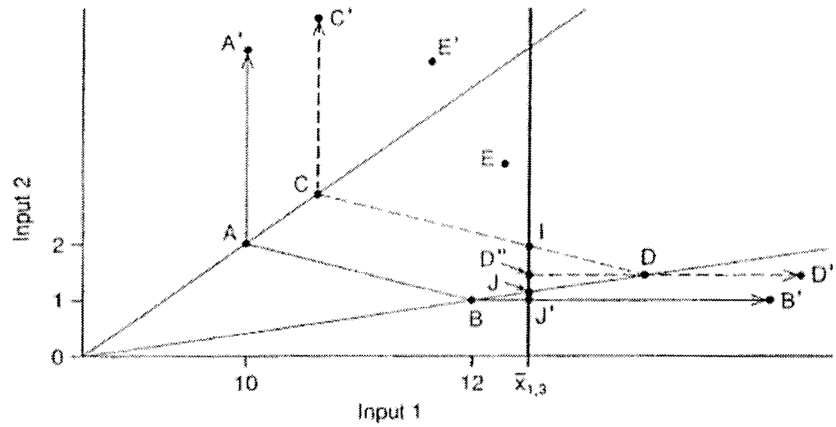

Figure 13-4. Lower Bound where $1 < \phi^* \leq \frac{1}{\theta}$

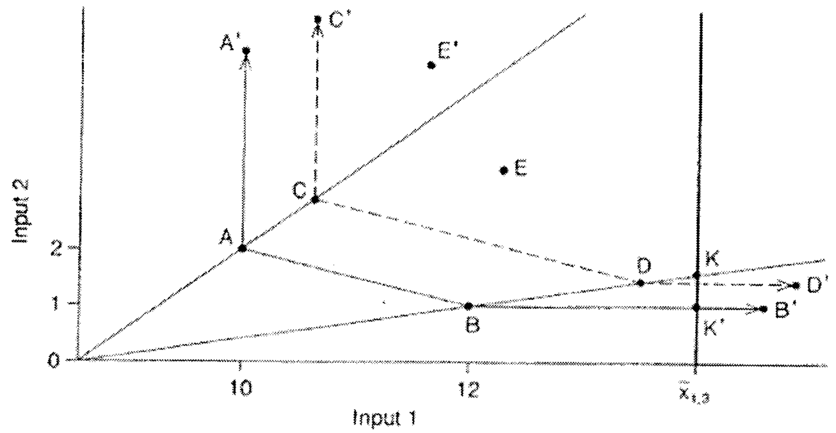

Figure 13-5. Lower Bounds where $\phi^* > \frac{1}{\theta}$

Clearly, any point on HD will suffice as a target input point.

In the case that lower bounds do not provide for access to the *frontier proper*, as is the situation described by Figures 13-4 and 13-5, then projection can take place *only* on to an *extension* of the frontier. In the example shown here, projection will be onto the extension BB'. In Figure 13-4 the target input would be D, and its projection would be back to J, then down to J'. Point D has a θ value of $\bar{\theta}$. As above, any point on ID other than D will have a θ value less than $\bar{\theta}$.

Figure 13-5 is a situation where θ will be strictly less than $\overline{\theta}$ for any feasible point. In this simple example, K will be the target point and its projection will be to K .

To formalize these ideas, let ξ denote the set of efficient (and properly enveloped) DMUs. We are now interested in those members ξ_b of ξ that are *input boundary points.* A DMU j_o is an input boundary point if there exists an input i_{j_o} such that $x_{i_{j_o}j_o} \leq x_{i_{j_o}j}$ for all $j = 1,\ldots,n$. That is, the minimum of the $x_{i_{j_o}j}$ across all DMUs j occurs in DMU j_o. Frontier extensions can emanate only from boundary points. Note in Figures 13-3, 13-4, 13-5 that both A and B are boundary points. That is, the minimum amount of input 1 is consumed by point A; the minimum of input 2 is consumed by point B. Thus, $\xi_b \equiv \xi = \{A, B\}$ for this example. The lower limit $\overline{x}_{13}$ on input 1 exempts point A from consideration, however, as a point from which a boundary extension can emanate. Therefore, only point B is of interest in terms of generating a surface for projections. In general, let $\overline{\xi}_b$ denote the subset of boundary points in ξ_b from which frontier extensions can emanate. Specifically, $\overline{\xi}_b$ consists of those DMUs j_o wherein the minimum component $x_{i_{j_o}j_o} \geq \overline{x}_{i_{j_o}n+1}$.

To derive a target input point, simply *expand* each member of $\overline{\xi}_b$ out into the feasible region (defined by Ψ). Specifically, solve the $\overline{J}$ linear programming problems:

$$\begin{aligned}&\min \phi_j\\&\text{subject to}\\&\phi_j x_{ij} \geq \overline{x}_{in+1}, i = 1,\ldots,I \qquad (13.9)\\&\phi_j \geq 0,\end{aligned}$$

where $\overline{J} = \{j \mid j\varepsilon\overline{\xi}_b\}$. Let ϕ_j^* denote the solution to (13.9) for a particular $j\varepsilon\overline{J}$. Define $\phi^* = \min_{j\varepsilon\overline{J}}\{\phi_j^*\}$, and let j_o be such that $\phi^* = \phi_{jo}^*$. If $\phi^* = 1$, then DMU j is feasible to Ψ and a portion of the efficient frontier is expanded as in Figure 13-3. In this case, solve (13.3) with (13.8) imposed to derive target inputs. If $1 < \phi^* \leq \frac{1}{\overline{\theta}}$, we have the situation described by Figure 13-4, and the target input point is given by $(x_{in+1},\ldots,x_{In+1})$ where $x_{in+1} = x_{ij_o}/\overline{\theta}$. If $\phi^* > \frac{1}{\overline{\theta}}$ as described by Figure 13-5, the target input point is given by $x_{in+1} = \phi^* x_{ij_o}$. In this case the optimal $\theta^* = \frac{1}{\phi^*}$ is strictly less than $\overline{\theta}$.

13.4. NONDISCRETIONARY VARIABLES

The discussion of the previous two sections centres around deriving a set of target inputs designed to achieve a given level of performance $\overline{\theta}$ if such a level is feasible. In the case where restrictive lower bounds are imposed on inputs it may not be possible to reach the floor performance score $\bar{\theta}$.

In the presence of a *nondiscretionary* input i_o, there is no choice regarding positioning in the i_o dimension. Refer again to Figures 13-3, 13-4, and 13-5. If x_1 is a nondiscretionary variable then the target input point must be located directly on the vertical line out of $\bar{x}_{13}$. In the case of Figure 13-3, this point will be at H$'$ which is horizontally opposite H (the line H$'$ H is parallel to the horizontal axis), if an efficiency rating of $\bar{\theta}$ is to be achieved. Here, the projection is to point G on the frontier. The point H$'$ is derived in general by modifying problem (13.3). Specifically, replace the set of inputs I by two subsets I_d and I_{nd}, representing the discretionary and nondiscretionary inputs, respectively. The second set of constraints in (13.3) is replaced by two sets of restrictions

$$\theta x_{in+1} - \sum_{j=1}^{n} \lambda_j x_{ij} - s_i = 0, i = 1,\dots,I_d \tag{13.10}$$

$$x_{in+1} - \sum_{j=1}^{n} \lambda x_{ij} - s_i = 0, i = I_{d+1},\dots,I, \tag{13.11}$$

The first set of constraints in (13.3) is replaced by

$$\alpha - x_{in+1} \geq 0, i = 1,\dots,I_d \tag{13.12}$$

and the objective function becomes

$$\min \alpha + \varepsilon \sum_{j=1}^{I_d} s_i. \tag{13.13}$$

Note that θ is set to $\bar{\theta}$ in (13.10). Assume that the inputs are renumbered so that the first I_d are the discretionary inputs, and so on.

In the case that no portion of the efficient frontier is exposed, we proceed as in Section 13.3, except that when $0 < \phi^* \leq \frac{1}{\bar{\theta}}$, the expression $x_{in+1} = x_{ij_o} / \bar{\theta}$ holds only for the discretionary variables $i = 1,\dots,I_d$. The other inputs are fixed at their nondiscretionary positions. This produces the point D$'$ in Figure 13.4.

In the situation that $\phi^* > \frac{1}{\bar{\theta}}$, the target inputs are located at K as was the case previously.

13.4.1 Uncertainty in the Nondiscretionary Variable

The above assumes that the value $x_{i_o n+1}$ for a nondiscretionary variable can be explicitly determined (for example, the demographic makeup of the population). If this value is *under estimated*, (for example, if the true value is $\hat{x}_{1n+1}$) however, then the use of $\theta = \bar{\theta}$ to choose the values of the discretionary variables x_{in+1} will not yield a projection on to the frontier (Figure 13-6). In this illustration where the projection is only to G$'$, it is clear that a smaller value of θ then $\overline{\theta}$ is needed in order to project the selected x_{in+1} onto the frontier. Obviously, if the value of x_{i_on+1} is over estimated, then the projection of the x_{in+1} using $\theta = \overline{\theta}$ will go below the frontier. This means only that larger values of the x_{in+1} could have been used. That is, the chosen x_{in+1} actually provides a higher efficiency then desired.

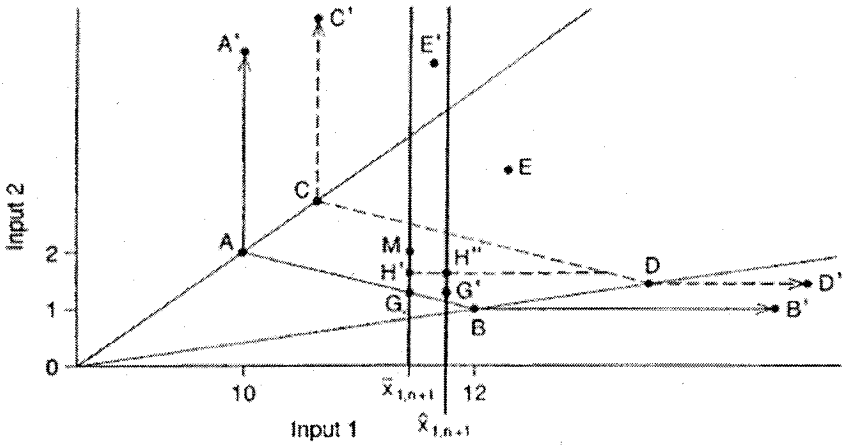

Figure 13-6. Under Estimation of Value of the Nondiscretionary Variable

To provide for the eventuality that the chosen value of the nondiscretionary variable may be under estimated, the following logic could be used: choose a set of target inputs for the discretionary variables such that if (1) the estimated value $\overline{x}_{in+1}$ of x_{i_on+1} occurs, then an efficiency of $\theta = 1$ results, but if (2) a higher value $\overline{x}_{i_on+1} + \beta$ of x_{i_on+1} occurs, the resulting efficiency $\theta \geq \overline{\theta}$. We propose solving the following linear problem in the case of a single discretionary variable:

$$\max \beta - \varepsilon(\sum_{i \neq i_o} x_{in+1})$$

subject to

$$x_{in+1} - \sum_{j=1}^{n} \lambda_j^1 x_{ij} = 0, i \neq i_o$$

$$\overline{x}_{i_on+1} - \sum_{j=1}^{n} \lambda_j^1 x_{i_oj} = 0$$

$$y_{rn+1} - \sum_{j=1}^{n} \lambda_j^1 y_{rj} \leq 0, \forall r$$

$$\bar{\theta} x_{in+1} - \sum_{j=1}^{n} \lambda_j^2 x_{ij} = 0, i \neq i_o$$

$$\bar{x}_{i_o n+1}^1 + \beta - \sum_{j=1}^{n} \lambda_j^2 x_{i_o j} = 0 \tag{13.13}$$

$$y_{rn+1} - \sum \lambda_j^2 y_{rj} \leq 0, \forall r$$

$$\sum_{j=1}^{n} \lambda_j^1 = 1$$

$$\sum_{j=1}^{n} \lambda_j^2 = 1$$

$$x_{in+1}, \lambda_j^1, \lambda_j^2, \beta \geq 0, \forall i, j.$$

In this problem we determine the maximum amount β by which $x_{i_o n+1}$ may increase from the estimated point $\bar{x}_{i_o n+1}$. Essentially, we are allowing for the maximal flexibility in the outcome of the uncontrollable variable $x_{i_o n+1}$. To ensure that a *proper facet* of the frontier is chosen in the selection of each of the two sets of multipliers $\{\lambda_j^1\}$ and $\{\lambda_j^2\}$, the term $-\varepsilon(\sum_{i \neq i_o} x_{in+1})$ is appended to the objective function. This forces the x_{in+1} as low as possible. In the figure, this would mean choosing x_{2n+1} at a point H$'$ as opposed to some higher point M. Of course, in this case since there is only one facet on the frontier, then such a point M would not actually arise, but could if other factors were present.

In the case of multiple nondiscretionary variables, there is no clear definition of maximal flexibility. Potential objectives might be to maximize the total deviation, i.e.

$$\max \sum_{i=I_{d+1}}^{I} \beta_i \tag{13.15}$$

or maximize the minimum deviation, i.e.,

$$\max \gamma \tag{13.16}$$

subject to

$$\gamma - \beta_i \leq 0, i = I_{d+1}, \ldots, I, \tag{13.17}$$

and subject to the other constraints in (13.13).

13.5. CONCLUSIONS

In this chapter we have examined the problem of setting input targets for new facilities when outputs are assumed to be known values. This problem is common in many retail settings where site location issues are involved. Both discretionary and nondiscretionary inputs are examined. All of the arguments here, regarding input targets, apply equally to the case where target outputs are at issue. In cases where standard size facilities are to be built, for example chain stores of a certain configuration, the required target market for the product mix becomes the focus of management. It is the output oriented model that would be the pertinent structure in this case. The development is straightforward, and is therefore omitted.

REFERENCES

1. Banker, R.D., A. Charnes and W.W. Cooper (1984), Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science* 30(9), 1078-1092.
2. Charnes, A., W.W. Cooper and E. L. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2(6), 429-444.
3. Cook, W.D., Y. Roll and A. Kazakov (1990), A DEA model for measuring the relative efficiency of highway maintenance patrols, *INFOR* 28(2), 113-124.
4. Morey, R., A.Y. Lewin and T. Cook (1982), Evaluating the administrative efficiency of courts, *Omega* 10(4), 401-411.
5. Seiford, L. (1996), Data envelopment analysis: the evolution of the state of the art, *Journal of Productivity Analysis* 7, 99-137.
6. Sherman, H.D. (1984), Hospital efficiency measurement and evaluation: empirical test of a new technique, *Med Care* 22(10), 922-938.
7. Sherman, H.D. and F. Gold (1985), Bank branch operating efficiency: evaluation with data envelopment analysis, *Journal of Banking and Finance* 9(2), 297-315.

This chapter is based upon Cook, W.D., R.H. Green and A. Kazakov, (1998), Setting performance targets for new decision making units in DEA, INFOR, 36(3), 177-188, with permission from University of Toronto Press.

Chapter 14

AGGREGATING PREFERENCE RANKINGS

14.1. INTRODUCTION

In a preferential election each voter selects a subset of k candidates from a ballot of m choices, and rank orders these k candidates from most to least preferred. Such a voting format is common in municipal elections, where a number of candidates are required to fill various positions. This same structure appears in other prioritization settings, such as the ranking of projects, products in a consumer survey, etc. We shall utilize the preferential voting example for discussion purposes throughout the chapter.

A problem of interest for over 200 years has to do with the *aggregation* of votes from preferential ballots. Borda (1781) proposed the "Method of Marks" as a means of deriving a consensus of opinions. This method amounts to determining the average of the ranks assigned by voters to each candidate, with the winning candidate being the one with the lowest average. An equivalent version of this model was later presented by Kendall (1962). Cook and Seiford (1982) have extended the Kendall model using an l^2 distance approach. Other distance based models have been advanced by Armstrong et al. (1977), Blin (1976), Cook and Seiford (1978), Cook and Kress (1984), Kemeny and Snell (1962), and others. In these distance models the voter ranks *all* of the alternatives or candidates.

Other models for aggregating preferential votes have arisen from parliamentary settings. The so-called American system, English system and West Australian system are examples of such models, as discussed in Keesey (1974). In some of these models the winning (first place) candidate is determined purely in terms of the maximum number v_{i1} of the first place votes. Here, v_{ij} denotes the number of jth place votes earned by candidate i. *Ties* are often broken (among candidates with the same number of first place votes) by utilizing 2nd place votes, 3rd place votes, etc. Under the American system, for example, the nominee having the least number of first choices is dropped. Those ballots on which that candidate was ranked first are now destroyed. The second choice votes received by the remaining candidates are now added to the first choices received by each. A nominee is declared elected if the total of his first and second place votes constitutes a majority – otherwise the candidate with the fewest number of first and second place votes is dropped, and the process is repeated by bringing third place votes into consideration.

Existing preferential voting models are arguably deficient in that they fail to provide a *fair overall* assessment of a candidate's standing --- a composite or aggregate evaluation of his first place, second place, third place standings, etc. The problem is one of how to combine the jth place votes, for $j = 1,2,...k$, in some rational manner.

In Section 14.2 a model is presented which utilized a *composite index* $z_i = \sum_{j=1}^{k} w_j v_{ij}$ of the jth place standings of candidate i. The objective of the model is to derive multipliers w_j (the level of importance attached to jth place votes) which will accord a candidate a *fair* assessment of his standing. This is accomplished by allowing for flexibility in the assignment of weights from one candidate to another. Since this process may result in several candidates being tied for first place, the model also provides for maximum *discrimination* among such candidates. This *robustness* property is a principal feature of the model.

The procedure used to derive the w_j, hence the index z_i, is analogous to the DEA method of Charnes, Cooper and Rhodes (1978). The usual DEA model possesses two important characteristics which could prove undesirable in the preference ranking setting. First, it is necessary to solve k problems; one for each decision making unit. Second, a number of units will commonly end up being tied for first place, i.e. will be on the efficient frontier. In Section 14.3 it is shown that the special properties of our composite index model eliminates both of these characteristics. Specifically, it is shown that only 1 problem, not k problems, needs to be solved to determine a winning candidate. Moreover, it is demonstrated that in most cases only one candidate will end up in the first place. Hence, ties for first place will generally be broken, and an ordinal ranking of the k best candidates emerges. It is also shown that a certain special case of this model is equivalent to Borda's method of marks for deriving a consensus among a set of ordinal rankings.

Examples and geometric interpretations are provided.

14.2. A FAIR MODEL FOR AGGREGATING PREFERENTIAL VOTES

In the preferential voting framework each candidate $i = 1, 2, \ldots, m$ receives some number v_{i1} of first place votes, v_{i2} of second place votes, ..., v_{ik} of kth place votes. The problem is to utilize these votes in a reasonable manner to obtain an overall *desirability index* z_i for each candidate.

For any given set of weights or multipliers w_j, we define the desirability index for candidate i by

$$z_i = \sum_{j=1}^{k} w_j v_{ij}. \tag{14.1}$$

Clearly, any preset values w_j are likely to favor some candidates while discriminating against others. In Borda's model, for example, the w_j are the numbers, $1,2,3,\ldots,n$. Moreover, the weights are the same for each candidate. What is required is a set of multipliers which provides the *fairest* possible treatment for each candidate. Specifically, we wish to determine a set of multipliers $\{w_j^{i_o}\}$ for candidate i_o which maximizes z_{i_o}. To achieve this, we solve the problem

$$z_{i_o}^* = \max \sum_{j=1}^{k} w_j v_{i_o j}$$

$$\text{subject to}$$

$$\{w_j\} \varepsilon \Phi, \tag{14.2}$$

where Φ is some subset of $\Re^k$.

In the context of preferential voting, the *feasibility* set Φ should be characterized by at least two types of constraints. First, z_{i_o} should be bounded above, i.e. $z_{i_o} \le \theta$ for some θ. Without such a parameter θ, problem (14.2) would be unbounded. Moreover, it is necessary to define some *best attainable* performance level (θ) that any candidate can achieve. As a convention, we set $\theta = 1$ or 100%. As will become apparent, the final rank ordering of the candidates is independent of the choice of θ.

The second set of constraints has to do with the priority attached to the jth versus $(j+1)^{st}$ place votes. It is clear that any reasonable aggregation scheme should be constrained by $w_j > w_{j+1}$. More generally, we define a function $d(j,\varepsilon): d: N\ X\ R^+ \to R^+$, i.e. $d(j,\varepsilon)$ is a non-negative function defined on the Cartesian product of the space of positive integers $N \equiv \{1,2,\ldots,K\}$ and the positive real line R^+. Moreover, $d(j,\varepsilon)$ is restricted to be non-decreasing in ε. We impose the constraints $w_j - w_{j+1} \ge d(j,\varepsilon)$. This lower limit $d(j,\varepsilon)$ on the gap between the importance attached to the jth versus $(j+1)$st place standing is referred to as the *discrimination intensity function*. The parameter ε is called the *discriminating factor.*

Problem (14.2) then becomes:

$$z_{i_o}(\varepsilon) = \max \sum_{j=1}^{k} w_j v_{i_o j}$$

subject to

$$\sum_{j=1}^{k} w_j v_{ij} \le 1, \qquad i = 1,2,...m,$$

$$w_j - w_{j+1} \ge d(j,\varepsilon), \qquad j = 1,2,...,k-1 \qquad (14.3)$$

$$w_k \ge d(k,\varepsilon).$$

This problem is now solved for each candidate $i_o = 1,2,...,m$.

Note that by defining a single "input" variable w_I and input data quantities $U_{iI} = 1,\ i = 1,...,m,$ problem (14.3) is equivalent to the well known DEA-AR model. See Thompson et al (1986, 1989). The constraints $w_j - w_{j+1} \ge (d,\varepsilon)$ represent the assurance region (AR).

To illustrate this problem, consider the case of four candidates, where two of these are to be elected from a preferential ballot. Let the 1st and 2nd place standings be given as follows:

		# Votes
Candidate i	v_{i1}	v_{i2}
1	6	8
2	4	11
3	8	2
4	3	0

Figure 14-1 shows the positioning of the four candidates in the $(v_{.1}, v_{.2})$ space. These are referred to as R_1, R_2, R_3, R_4.

If problem (14.3) is now solved for each of the four candidates, the constraint space appears as in Figure 14.2. In this diagram we have set $d(j,\varepsilon) \equiv 0$. Candidate #1 has an optimum either at point $A\left(w_1 = \frac{3}{34}, w_2 = \frac{3}{34}\right)$ or point $B\left(w_1 = \frac{3}{26}, w_2 = \frac{1}{26}\right)$. In any event, the rating for candidate #1 is $z_1^* = 1.0$. Similarly, candidate #2 has its optimum at the point A where $z_2^* = 1.0$, and candidate #3 at point B where $z_3^* = 1.0$. Candidate #4 has its optimum at point C, where $z_4^* = \frac{3}{8}$.

Referring back to Figure 14-1, if the points representing candidates #1 and #2 (i.e. $R1$ & $R2$) are joined, the resulting line segment has a slope given by the ratio of the w_1, w_2 coordinates of point A in Figure 14.2. We refer to the set of line segments joining points $R2 - R1 - R3$ as the *desirability frontier*, in that candidates i on this boundary have achieved the highest attainable index $z_i = 1.0$. In fact, the desirability rating for candidate #4 is given by the ratio of the line segment or range from the origin 0 to R4 to the line segment 0 to R4'.

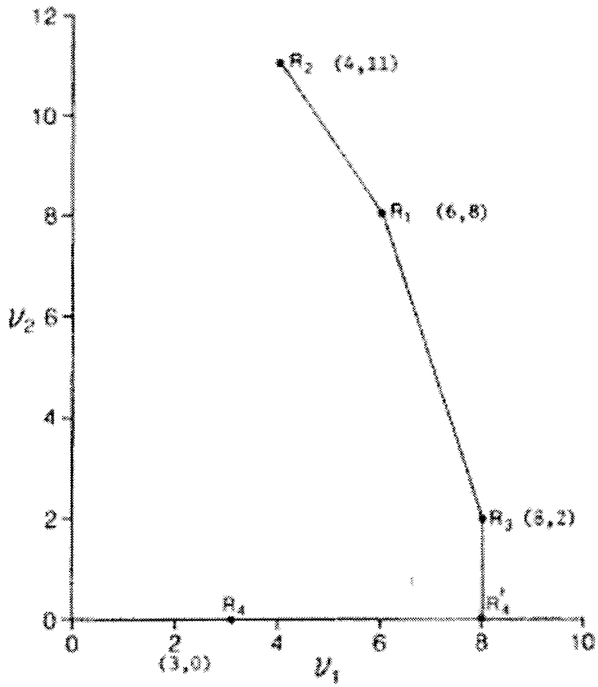

Figure 14-1. Desirability Frontier

Two important points must be emphasized here. First, different multipliers (w_1, w_2) were used for the different candidates. Point A in Figure 14-2 defines the importance attached to first and second place votes in evaluating the standing of candidate #2. Point B is used to evaluate candidate #3. Points A and B are both optimal for candidate #1, and C is optimal for candidate #4.

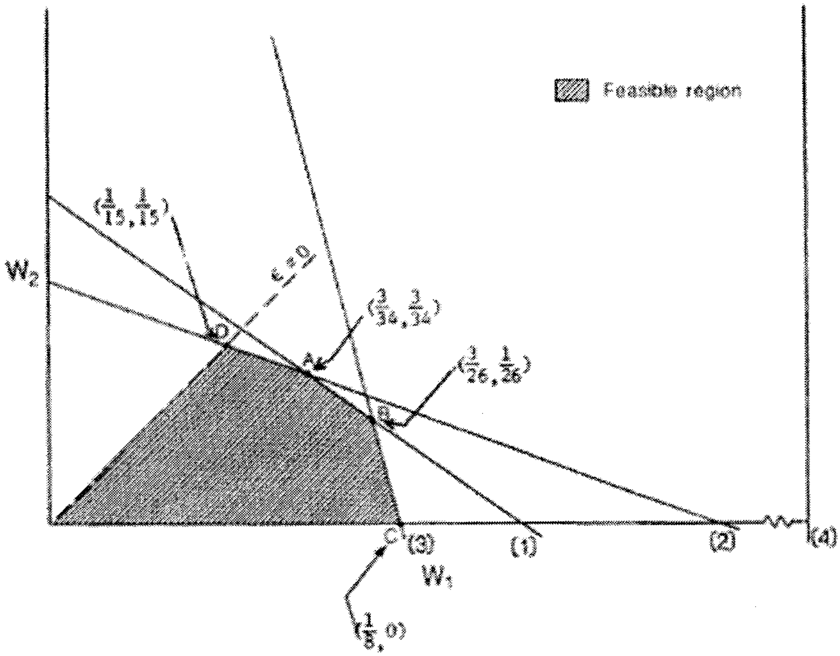

Figure 14-2. Feasible Weight Space for ε = 0

Second, if $d(j,\varepsilon) = \varepsilon$ and ε increases from its 0-level, the optimal points for certain candidates can change. Figure 14-3 shows the shape of the feasible region when ε = 1/27. Here, the optimum for candidate #2, for example, is located at E. The resulting different coefficients of E, versus those of A, lead to a redefinition of the desirability frontier. See line segment $R1 - R2'$ in Figure 14-4. Now, candidate #2 no longer has an index of $z_2^* = 1$, meaning that the number of possible first place candidates has been reduced from three to two.

Clearly, the value of the discriminating factor ε influences the ranking of the candidates, in the sense that it discriminates among candidates on the frontier. In the section to follow we examine the impact of this factor.

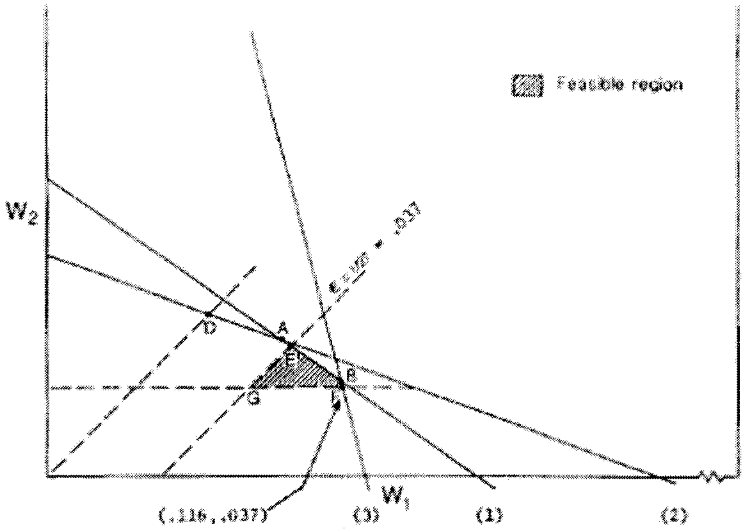

Figure 14-3. Feasible Weight Space for ε = 1/27

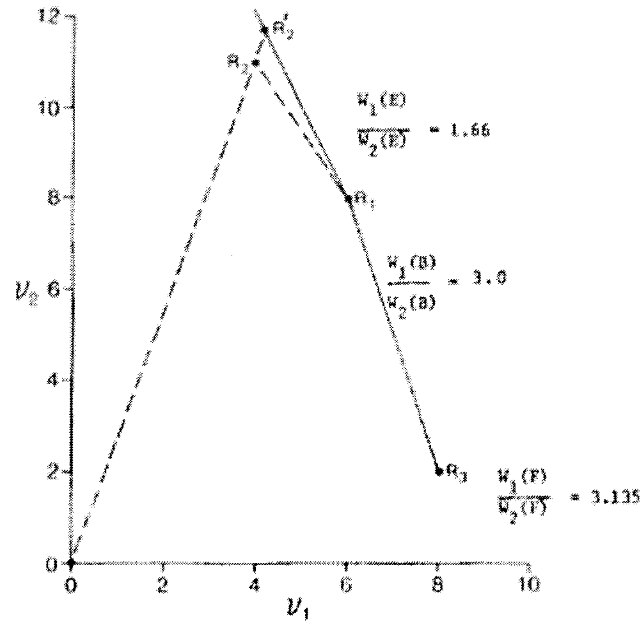

Figure 14-4. Adjusted Desirability Frontier

14.3. A MODEL FOR RANKING THE CANDIDATES

In model (14.3), for any given value of ε, one or more candidates i will achieve the maximum attainable desirability index $z_i(\varepsilon)=1$. In this section we show that there is a maximum value for ε for which this problem has a solution, and demonstrate that the derivation of this value can lead to an ordinal ranking of the candidates. First a number of properties of (14.3) and $z_i(\varepsilon)$ are presented.

Property 14.1:
For any i_o, $z_{i_o}(\varepsilon)$ is a monotonic non-increasing function of ε.
Proof:
From the monotonicity of $d(j,\varepsilon)$, it is clear that the feasibility region in (14.2, 14.3), for a given ε, contains the corresponding region for ε', if $\varepsilon' > \varepsilon$. Therefore, the objective function cannot increase.

Q.E.D.

Corollary 14.1:
If $z_{i_o}(\varepsilon_1)=1$ *then* $z_{i_o}(\varepsilon)=1$ *for all* $\varepsilon \le \varepsilon_1$.

Proof:
This follows from Property 14.1 and the i_o-th constraint in (14.3). For a given function $d(j,\varepsilon)$, define $I_1=\{i; z_i(0)=1\}$. We assume that $d(j,\varepsilon)$ is defined such that $I_1 \ne \varnothing$. For any $i \varepsilon I_1$ let ε_i^* be the largest value of ε such that $z_i(\varepsilon)=1$, and let $\varepsilon_{\max}^* = \mathrm{Max}_i\{\varepsilon_i^*\}$. Clearly $\varepsilon_{\max}^*$ may be ∞, for example when $d(j,\varepsilon)=0$ for all j and ε.

Q.E.D.

Property 14.2:
$\varepsilon_{\max}^*$ is the largest value of ε for which there exists a feasible solution for (14.3).

Proof:
Without loss of generality we can assume that $\varepsilon_{\max}^* < \infty$. Let $\varepsilon > \varepsilon_{\max}^*$ and suppose that there exists a feasible solution $w=(w_1,\ldots,w_k)$ for (14.3). Then, by the definition of $\varepsilon_{\max}^*$ we must have that $z_i(\varepsilon)<1$ *for all* $i,\ldots,m$. The dual of (14.3) is

$$\min \sum_{i=1}^{m} x_i - \sum_{j=1}^{k} d(j,\varepsilon) y_j$$

subject to

$$\sum_{i=1}^{m} x_i v_{i1} - y_1 \geq v_{i_o 1} \tag{14.4}$$

$$\sum_{i=1}^{m} x_i v_{ij} + y_{j-1} - y_j \geq v_{i_o j}, \qquad j = 2...k,$$

$$x_i, y_i \geq 0.$$

From complementary slackness it follows that at optimality, $x_i^* = ... = x_m^* = 0$. This implies that at least one of the constraints in (14.4) is violated, in contradiction. We conclude that the problem of determining the desirability index for each alternative applies only for ε in the closed interval $\left[0, \varepsilon_{\max}^*\right]$, which is the *feasible range* of ε.

Q.E.D.

Theorem 14.1:

There exists an alternative i_o such that $z_{i_o}(\varepsilon) \geq z_i(\varepsilon)$, $\forall\ i = 1, ..., m$, $\forall \varepsilon$ in [0, $\varepsilon_{\max}^*$].

Proof:

Let alternative i_o be such that $\varepsilon_{i_o}^* = \varepsilon_{\max}^*$. Hence, $z_{i_o}(\varepsilon) = 1$, for all values of ε in the feasible range. Pick any alternative i such that $z_i(\varepsilon) < 1$ for $\varepsilon_i^* < \varepsilon \leq \varepsilon_{\max}^*$. We conclude that there exists a nonempty set S of alternatives such that for *all* feasible values of the discriminating factors ε, the alternatives i_o in S are superior, in terms of the desirability index, to all other alternatives.

Q.E.D.

Consider the problem

$$\max \varepsilon \tag{14.5a}$$

subject to

$$\sum_{j=1}^{k} w_j v_{ij} \leq 1, \qquad i = 1...m, \tag{14.5b}$$

$$w_j - w_{j+1} - d(j,\varepsilon) \geq 0 \tag{14.5c}$$

$$w_k - d(k,\varepsilon) \geq 0 \tag{14.5d}$$

$$w_i, \varepsilon \geq 0. \tag{14.5e}$$

Property 14.3:

At optimality, at least one of the constraints in (14.5b) holds at equality, that is there exists a candidate i_o such that $\sum_{j=1}^{k} w_j v_{i_o j} = 1$.

Proof:
This follows directly from the proof of Property 14.2.

Q.E.D.

It follows from Property 14.3 and Theorem 14.1 that the solution to (14.5) yields all the winning candidate(s). Specifically, the candidates i for which $\sum_{j=1}^{k} w_j v_{ij} = 1$ at $\varepsilon_{\max}$, dominate all other candidates in that they possess the highest level of *robustness* in terms of the range of ε. For the case $d(j,\varepsilon) = \varepsilon$, in a random sample of 60 problems, each with 20 voters and 10 candidates, 93% of the cases yielded a *single* first place candidate.

Having derived the first-place candidate i_o, the i_oth constraint in (14.5b) can now be eliminated and (14.5) resolved. That candidate i_1 with $\sum w_j v_{i_1 j} = 1$ will hold second place, and the process is repeated.

The following theorem characterizes the nature of the w_j differentials vis-à-vis the discrimination intensity function.

Theorem 14.2:
In optimality, all the constraints in (14.5c) and (14.5d) hold as equalities.

Proof:
Let $I_o = \{i; \sum w_j^* v_{ij} = 1\}$ and, without loss of generality, assume that $v_{i1} > 0$ for all $i \in I_o$. Suppose that an optimal solution of (14.2) is such that $w_s - w_{s+1} - d(s,\varepsilon) = \delta > 0$, for some s. Define

$$w_j' = \begin{cases} w_j - \delta, & j = 1 \ldots s, \\ w_j, & j = s+1, \ldots, k. \end{cases}$$

Clearly,

$$\sum_{j=1}^{k} w_j' v_{ij} < \sum_{j=1}^{k} w_j v_{ij} \le 1 \quad \forall i \in I_o,$$

$$\sum_{j=1}^{k} w_j' v_{ij} \le \sum_{j=1}^{k} w_j v_{ij} < 1 \quad \forall i \notin I_o,$$

$$w_j' - w_{j+1}' - d(j,\varepsilon_{\max}^*) \ge 0, \text{ with } w_s' - w_{s+1}' - d(s,\varepsilon_{\max}^*) = 0.$$

It follows that $(w', \varepsilon_{\max}^*)$ is also an optimal solution for (14.2). For w', all the constraints in (14.5b) are strict. But, according to Property 14.3 this is impossible. Q.E.D.

Corollary 14.2:
Problem (14.5) is equivalent to

$\max \varepsilon$

subject to

$$\sum_{j=1}^{k}(\sum_{\ell=j}^{k} d(l,\varepsilon))v_{ij} \leq 1, \qquad i = 1...m, \tag{14.6}$$

$$\varepsilon \geq 0.$$

Proof:

From Theorem 14.2, the w_j can be recursively expressed as sums of the $d(l,\varepsilon)$. Q.E.D.

The ease with which problem (14.5) can be solved depends very much on the form of the $d(j,\varepsilon)$ function. If d is a linear function, for example, then (14.5) will be a linear programming problem. Such would be the case say where $d(j,\varepsilon) = a_j\varepsilon$, with $\{a_j\}$ being a set of scalars. Special cases of this would be: (i) a_j = constant; (ii) $\{a_j\}$ is a monotonic increasing sequence (larger gaps between w_j & w_{j+1} as j increases); (iii) $\{a_j\}$ is a decreasing sequence (smaller gaps as j increases).

An important subclass of discrimination intensity functions is that for which $d(j,\varepsilon) = g(j)h(\varepsilon)$, where $h(\varepsilon)$ is strictly monotonic increasing in ε. The linear examples of the previous paragraph fall into this category. For this subclass, problems (14.5) and (14.6) have closed-form solutions. Specifically, problem (14.6) becomes:

$$\max \varepsilon$$

subject to

$$h(\varepsilon)\sum_{j=1}^{k}(\sum_{l=j}^{k} g(l))v_{ij} \leq 1, \qquad \varepsilon \geq 0, \qquad \text{and}$$

$$\varepsilon^*_{\max} = \min_i h^{-1}[(\sum_{j=1}^{k}(\sum_{l=j}^{k} g(l))v_{ij})^{-1}]. \tag{14.7}$$

A winning candidate i is, therefore, the one for which this minimum is attained. Due to the monotonicity of h, it is sufficient to find the candidate i for which

$$\sum_{j=1}^{k}\sum_{l=j}^{k}(g(l))v_{ij} \tag{14.8}$$

is maximized.

Special cases:

(1) Exponentially decreasing intensity of discrimination values: $g(j) = e^{-\alpha j}$. The maximum of (14.8) then becomes:

$$\max_i \sum_{j=1}^{k}\left(\frac{1-e^{-\alpha(k-j+1)}}{e^{\alpha j}-1}\right)v_{ij.}$$

The candidate i, for which the above maximum is attained, is the winner.

(2) Constant intensity of discrimination values: $g(j) = 1$. In this case, the winner is a candidate with the maximum value of $\sum_{j=1}^{k}(k - j + 1)v_{ij}$.

This is, of course, the well-known Borda method of marks. Consequently, Borda's (and Kendall's 1962) model for deriving a consensus among a set of voter rankings is a special case of our model when $g(j) =$ a constant.

Example:

Consider the case of 20 voters, each of whom is asked to rank 4 out of 6 candidates on a ballot. Let the outcome from the vote be as shown in matrix V.

		Standing			
		1	2	3	4
	a	3	3	4	3
	b	4	5	5	2
V=	c	6	2	3	2
	d	6	2	2	6
	e	0	4	3	4
	f	1	4	3	3

For example, candidate "a" receives 3 first, 3 second, 4 third and 3 fourth-place votes. Specifically, $v_{a1} = 3$, $v_{a2} = 3$, $v_{a3} = 4$, $v_{a4} = 3$.

Case 1: $d(j,\varepsilon) = \varepsilon \Rightarrow b$ is the winner, $\varepsilon^*_{max} = 0.0233$. (This is the Borda/Kendall winner.)

Case 2: $d(j,\varepsilon) = \varepsilon / j \Rightarrow d$ is the winner, $\varepsilon^*_{max} = 0.0577$.

Case 3: $d(j,\varepsilon) = \varepsilon / j! \Rightarrow c$ is the winner, $\varepsilon^*_{max} = 0.0808$.

In Case 1 the discrimination intensity is evenly distributed among the rank positions; therefore the fact that candidate b had the largest number of votes for rank positions 1, 2 and 3 played a major role in determining him/her as the winner. In Case 3 on the other hand, the discrimination intensity function is exponentially decreasing in j, which implies that the first place vote has a relatively very large weight while the fourth position has very little weight in determining the winner. Although candidate d received more total votes (16) than candidate c (13), the fact that d's advantage was concentrated in the fourth position while in the third position

c had a slight advantage on d, (in positions 1 and 2 they had identical scores) plays a key role in determining c as a winner for the $d(j,\varepsilon) = \varepsilon / j!$ case.

14.4. CROSS EVALUATION

The above analysis, as provided by problem (14.5), might be criticized as running counter to the objective of providing the fairest possible treatment of each candidate. The principal conclusion is that for discrimination intensity functions of the form $d(j,\varepsilon) = \varepsilon$, this problem is equivalent to applying the Borda-Kendall count technique.

While in some situations it may be desirable to have a common (imposed) set of weights, clearly the flexibility to choose the most favorable standing for each candidate can be compromised. Specifically, adopting a starting point of fairness in evaluation inherent in (14.3) is compromised by a commitment to discrimination via (14.5).

Before proposing an alternative to the approach given above, we rewrite problem (14.3). Specifically, we make the change of variables

$$V_{qj} = \sum_{x=1}^{j} v_{qx},$$

and (14.3) becomes, by virtue of Theorem 14.2 and Corollary 14.2:

$$Z_{ii} = \text{Maximize } \sum_{j=1}^{k} W_{ij} V_{ij} \tag{14.9a}$$

Subject to:

$$Z_{iq} = \sum_{j=1}^{k} W_{ij} V_{qj} \leq 1 \qquad \text{for } q = 1,2...,m. \tag{14.9b}$$

The relationship between weights $w_{..}$ and $W_{..}$ is given by

$$w_{ij} = \sum_{x=1}^{j} W_{ix}$$

One interpretation of the Cook and Kress (CK) method is to examine it in terms of "who" is choosing the weights. While candidates in general do not choose their own weights in (14.3), it may appear that the *winner* does choose his/her own weights. The winner in (14.3) effectively does so by establishing a desirability of 1.0 with the weights:

$$w_{pj} = h(\varepsilon_{\max})(\sum_{x=j}^{k} g(x)) \qquad \text{where } h(\varepsilon_{\max}) = 1/\pi_p$$

The other candidates are then ranked using these weights. CK's procedure, in a sense, yields the winner and the winner's weights are used to rank order the other candidates. Thus, candidates are ranked according to how they are *cross-evaluated* (Sexton et al. 1986; Oral et al. 1991) by the *winner(s),* P.

In order to retain the essence of CK's approach, i.e. (14.9(a)-(b)), but still discriminate between candidates, we are now motivated to investigate how the idea of cross-evaluation, by *all* candidates, within this model can be used to arrive at an overall rating of each candidate.

When model (14.9) is solved for candidate i, as well as obtaining Z_{ii}, we are also provided with values Z_{iq}, which can be thought of as evaluations of q's desirability from i's point of view – within this modeling framework. The values obtained in a complete run of the model can be organized in a matrix Z in which the values across a row $i(Z_{i.})$ represent how candidate i *rates* all candidates and values down column $i(Z_{.i})$ represent how candidate i *is rated by* all candidates. Thus, the matrix can be regarded as the summary of a self- and peer-rating process in which on-diagonal elements represent self-ratings and off-diagonal elements represent peer-ratings. Such a matrix, for the example in Table 14-1, is shown in Table 14-1

Table 14-1. Votes achieved by candidates a-f

Candidate	1	2	3	4
a	3	3	4	3
b	4	5	5	2
c	6	2	3	2
d	6	2	2	6
e	0	4	3	4
f	1	4	3	3

Table 14-2. A cross-evaluation matrix for the example of Table 14-1

	a	b	c	d	e	f
a	0.813	1.0	0.813	1.0	0.688	0.688
b	0.667	1.0	0.786	0.714	0.35	0.475
c	0.5	0.667	1.0	0.95	0.0	0.167
d	0.5	0.667	0.813	1.0	0.0	0.167
e	0.813	1.0	0.813	1.0	0.688	0.688
f	0.813	1.0	0.813	1.0	0.688	0.688

The problem now is to arrive at an overall rating for each candidate, consistent with the self- and peer-ratings in Z, which can then be used to rank the candidates. Our first inclination is to follow Sexton et al. (1986) and regard the column averages of Z as suitable overall ratings. In essence, each candidate is being allowed equal right to interpret the voter's preferences, (i.e. all the candidates' standings) as manifested in Z, rather than just the winner(s) according to CK's approach. Thus, each candidate is being accorded a weight of 1/m in determining any candidate's overall rating.

Applying this idea to the cross-evaluation matrix in Table 14-1, the column averages and corresponding ranking is as follows:

d(.944) >b(.899) >c(.840) >a(.684) >f(.479) >e(.402).

However, consider the situation if there were two extra candidates, *g* and *h* say, each receiving one third-place vote. The cross-evaluation matrix for this situation is given in Table 14-1. (To reflect the fact that the electorate is finite, with twenty voters, we have deducted one third place vote from each of candidates, e and f.)

Table 14-3. A cross-evaluation matrix for the example of Table 14-1 after appending two extra candidates

	a	b	c	d	e	f	g&h
a	0.813	1.0	0.813	1.0	0.688	0.688	0.063
b	0.667	1.0	0.786	0.714	0.35	0.475	0.0
c	0.5	0.667	1.0	0.95	0.0	0.167	0.0
d	0.5	0.667	0.813	1.0	0.0	0.167	0.0
e	0.813	1.0	0.813	1.0	0.688	0.688	0.063
f	0.813	1.0	0.813	1.0	0.688	0.688	0.063
g&h	0.714	1.0	0.786	0.714	0.429	0.5	0.071

*A cross-evaluation matrix for the example of Table 14-1 after appending two extra candidates, g and h, who each receive one third-place vote. The standings of candidates e and f have each been reduced by one third-place vote to reflect the finite electorate of twenty voters.

It can be seen that the order of the candidates is now:

b(.917) >d(.877) >c(.826) a(.692) >f(.484) >e(.409) g=h(.041); there has been a reversal in the positions of candidates b and d consequent on the introduction of the two lowly rated candidates g and h. The 'principle of the independence of irrelevant alternatives' (e.g. Arrow, 1951), has been contravened. However, French (1986) suggests, echoing many other commentators, that this principle is

"*...arguably the most controversial assumption within social choice theory, not to say within decision theory...*".

For our own part, we must be concerned by a rank reversal provoked by a very lowly rated candidate, such as in the example above, but this concern would not necessarily extend to all contexts. In the turbulence following in the wake of the exit from the scene (or entry to the scene) of a *highly* rated candidate, we perhaps should not be surprised to see some change in the order of the candidates.

In order to mitigate the effect observed above, we may relax the assumption that each candidate be accorded a weight of 1/m in the establishment of the overall ratings. Instead, we suggest that each candidate applies a weight in proportion to his/her overall rating rather than uniformly

1/m i.e. a form of 'weighted voting' (Tideman, 1976). Thus if Θ is the row vector of final ratings, we now solve:

$$(1/\sum_{i=1}^{m}\theta_i)\ \Theta Z = \Theta \tag{14.10}$$

to obtain candidate i's overall rating, θ_i.

For the cross-evaluation matrices of Tables 14.3 and 14.4 we now obtain:

d(.928) >b(.852) >c(.847) >a(.643) >f(.411) >e(.310) and

d(.926) >b(.854) >c(.846) >a(.644) >f(.412) >e(.312) >g =h(.022)

respectively. The previous rank reversal does not now occur and, perhaps more important, the overall ratings are scarcely changed between the two situations.

In order to solve (14.10), we proceed iteratively as follows:

Step 1: $\Theta^{\text{new}} = (1/\sum_{i=1}^{m}\theta_i^{\text{old}})\Theta^{\text{old}}Z$ (14.11)

Step 2: $\Theta^{\text{old}} = \Theta^{\text{new}}$

Initially we set Θ^{old} as (1,1,...,1) so $1/\sum\theta_i^{\text{old}}$ is $1/m$, and the first iteration gives the vector of simple column averages of Z. This is then refined in subsequent iterations. The algorithm in (14.11) is essentially the "Power Method" for the principal eigenvalue/eigenvector of Z. While this method is not particularly efficient as a means of obtaining principal eigenvalue/eigenvectors of arbitrary matrices, it works very well for cross-evaluation matrices, converging to 3 decimal places in about five iterations.

The resemblance of (14.11) to the Power Method is not surprising, or course, since (14.10) can be rewritten as:

$$\Theta z = (\sum_{i=1}^{m}\theta_i)\Theta\ , \tag{14.12}$$

whereupon Θ can be seen as the left-hand eigenvector of Z, scaled to an eigenvalue of $\sum\theta_i$. Our proposal can now be seen to be reminiscent of Wei's rating method (Wei, 1952; but see Cook and Kress, 1992), and also Saaty's "Analytic Hierarchy Process" (Saaty, 1977; 1994), which both use the principal eigenvector of a matrix as representing an overall rating. Of course, the meanings of the matrices and the manner of their determination in these two cases, are somewhat different from our cross-evaluation matrix.

14.5. CONCLUSION

The problem of aggregating votes in a preferential election has been a subject of study for over 200 years. In this chapter a model is presented which aggregates votes into an overall index, in a way that allows each candidate to be assessed in a fair manner. The model amounts to determining, for each candidate i, the best set of weights w_j to apply to j th place standings v_{ij} for that candidate. We define a discrimination intensity function $d(j,\varepsilon)$ which specifies the minimum amount by which the multipliers w_j and w_{j+1} must differ. It is shown that for a certain general subclass of these functions, the winning candidate can be obtained by a closed-form expression. Furthermore, it is shown that for the special case in which all consecutive pairs of weights deviate by the same amount $(d(k,\varepsilon)=d(j,\varepsilon))$, our model is equivalent to the well-known models of Borda and Kendall.

REFERENCES

1. Armstrong, R.D. and W.D. Cook (1977), A committee ranking problem using ordinal utilities, *Research Report CCS 298,* Center for Cybernetic Studies, University of Texas at Austin.
2. Arrow, K.J. (1951), Social choice and individual values, Yale University Press, New Haven and London.
3. Blin, J.M. (1976), A linear assignment formulation of the multiattribute decision problem, *Rev. Francaise d'Automatique, Informatique et Recherche Oper.,* 10, 21-32.
4. Borda, J.C. (1781), Memoire sur les elections au scrutin, *Histoire de l'Acad. Royale Sci.,* Paris.
5. Charnes, A., W.W. Cooper and E. Rhodes (1978), Measuring efficiency of decision making units, *European Journal of Operational Research* 2, 429-444.
6. Cook, W.D., and M. Kress (1984), Relationships between l^1 metrics on linear ranking spaces, *SIAM Journal of Applied Mathematics* 44, 209-220.
7. Cook, W.D., and M. Kress (1992), Ordinal information and preference structures: Decision models and applications, Prentice Hall, New Jersey.
8. Cook, W.D., and L.M. Seiford (1982), The Borda-Kendall consensus method for priority ranking problems, *Management Science* 28, 621-637.
9. Cook, W.D., and L.M. Seiford (1978), Priority ranking and consensus formation, *Management Science*, 24, 1721-1732.
10. French, S. (1986), Decision Theory, Ellis Horwood, Chichester.

11. Keesey, R.E., (1974), Modern Parliamentary Procedure, Houghlin Mifflin Co., Boston.
12. Kemeny, J.G. and L.J. Snell (1962), Preference ranking: An axiomatic approach, in *Mathematical Models in the Social Sciences*, Ginn, New York, 9-23.
13. Kendall, M. (1962), Rank Correlation Methods, 3rd ed. Hafner, New York.
14. Oral, M., O. Kettani and P. Lang (1991), A methodology for collective evaluation and selection of industrial R&D projects, *Management Science* 37(7), 871-883.
15. Saaty, T.L. (1977), A scaling method for priorities in hierarchical structures, *Journal of Mathematical Psychology* 15, 234-281.
16. Saaty, T.L. (1994), Highlights and critical points in the theory and application of the analytic hierarchy process, *European Journal of Operational Research* 74(3), 426-447.
17. Sexton, T.R., R.H. Silkman and A.J. Hogan (1986), Data envelopment analysis: critique and extensions, in R.H. Silkman, ed. *Measuring efficiency: an assessment of data envelopment analysis,* Jossey-Bass, San Francisco.
18. Tideman, T.N. (1976), The capabilities of voting rules in the absence of coalitions, *Policy and Politics* 4, 23-44.
19. Thompson, R.G., F.D. Singleton, Jr., R.M. Thrall and B.A. Smith (1986), Comparative site evaluations for locating a high energy lab in Texas, *Interfaces*, 1380-1395.
20. Thompson, R.G., L.N. Langemeiar, C.T. Lee and R.M. Thrall (1989), The measurement of productive efficiency with an application to Kansas Royland Wheat Farming, Jesse H. Jones Graduate School of Administration working paper, #65.
21. Wei, T.H. (1952), The Algebraic Foundation of Ranking Theory, Unpublished thesis, Cambridge University.

This chapter is based upon Cook, W.D. and M. Kress, (1990), A data envelopment model for aggregating preference rankings, Management Science, 36(11), 1302-1310, with permission from the Institute of Management Sciences, and Doyle, J.R., W.D. Cook, and R.H. Green, (1996), Preference voting and project ranking using DEA and cross evaluation, European Journal of Operational Research, 90(3), 461-472, with permission from Elsiever Publishing.

Chapter 15

RANKING PLAYERS IN ROUND ROBIN TOURNAMENTS

15.1. INTRODUCTION

In a round robin tournament, each of n players competes with every other player exactly once, with each match resulting in a decision (no draws). The problem of rank ordering the players based upon the results from the competition has been studied by numerous authors. These include Ali et. al. (1986), Goddard (1983), Kendall (1962), Wei (1952), Cook et al (1988a; 1988b; 1988c; 1990; 1992), Moon (1968), and others. The player ranking problem is very often approached in a 2-stage fashion. In stage one, a player's performance is used to arrive at a *rating* for that player. Stage two then takes that set of ratings and creates an ordinal *ranking*. The necessity of stage two arises due to the presence of ties in the ratings of stage one, and, therefore, is concerned primarily with algorithms for breaking such ties (see e.g. Goddard (1983)). It is the stage one problem with which we concern ourselves herein.

Tournament ranking theory has been applied in many areas including the problem of prioritizing transportation projects. In Cook et. al. (1988) such a model has been developed within the Ministry of Transportation, Ontario, Canada. In this context, projects are rated across multiple dimensions, and evaluated using a concordance model of the ELECTRE type (Roy 1968). The result of this analysis is a binary preference matrix with a structure similar to that of tournament. Since the ultimate desire is to rank order the projects based on this binary matrix, a tournament algorithm is applied at this stage.

In section 2 of this chapter, we begin by examining strict tournaments, that is where matches between players always result in a win. We briefly review some of the existing player rating models, and explore possible weaknesses which these models may exhibit. From these observations, we then establish a set of criteria which should be adhered to in developing a rating based on the outcomes of a tournament. In particular, these criteria call for a consideration of *strength* of a player j in accounting for the worth of a win of i over j. Specifically, we consider not only the immediate or 1st generation wins, but also 2nd generation wins, 3rd generation wins, ..., etc. A tournament rating model is then presented which takes account of the strength factor. We briefly discuss the geometric interpretation of the model in terms of the Data Envelopment Analysis (DEA) constructs of Charnes, Cooper & Rhodes (1978) and others. An illustrative example is included. Finally, we present an enumeration algorithm for obtaining m^{th} generation scores.

In section 3 we extend these ideas to the case of weak tournaments. Section 4 examines the ranking of players in multiple tournaments, and Section 5 looks at cases where only partial tournaments may exist. Conclusions are presented in Section 6.

15.2. A MODEL FOR RATING PLAYERS

Existing Rating Methods

A number of approaches have been suggested in the literature for rating n players, using as the basis for the ratings, the adjacency matrix $A = (a_{ij})$. Recall that $a_{ij} = 1$ if player i defeats player j, and $a_{ij} = 0$ otherwise. The concept in these approaches is to develop a set of scores from the elements of A. We briefly review some of these methods:

1. *Row-Sum or Kendall Scores* (Kendall (1962)). This approach amounts to computing the sum of the elements $a_{i.}$ in each row i of A, and using this as the rating for that player. This technique counts only direct wins, and makes no attempt to measure the *strength* of players.

2. *Row-Sum Scores of a Higher Power A^P of A* (Wei (1952)). Here, a win against a stronger player is awarded a higher score than one against a weaker player, that is, the strength of the opponent is a major factor in determining the score of the player. However, the number of wins which a player achieves is not directly taken into account.

3. *Row-Sum of p-Connectivity Matrices* (Goddard (1983)). The r^{th} stage p-connectivity matrix (p_{ij}^r) is defined by

$p_{ij}^{r} = \{\sum_{k=1}^{n} p_{ik}^{r-1} + \sum_{k=1}^{n} p_{jk}^{r-1}\}a_{ij}.$

The concept here is to continue to generate higher powers P^{r} until the row sums of P^{r} are no longer tied. This method is somewhat between the previous two methods, but also fails to take account of player strength in a reasonable manner.

Criteria for Rating Players

Each of the aforementioned methods exhibits both strengths and weaknesses vis-a-vis the attainment of a reasonable rating of the players. The following is a proposed set of criteria which any player rating method should adhere to:

- The *direct scores* or wins of each player should be taken into account.
- Ratings should be based on the *strength* of the defeated opponents.
- The method should rate each player relative to the others in the *fairest* or best possible way (give each player his "best shot".)

The Model

The model we propose for rating players is based on determining a set of weights $0 \leq W_i \leq 1$ such that the above criteria are observed. The approach looks at the entire range of wins of a player i over players j, not only in the immediate sense (i beats j), but in the more remote or m^{th} generation senses (i beats k who beats j). We give the following definitions.

Definition 15.1: The *digraph* $G(T)$ of tournament T is that graph whose set of nodes V consists of the n players, and whose edges E represent the outcomes (wins and losses) of the competition.

Definition 15.2(a): The m^{th} Generation Score (MGS) of player i is the number of nodes j for which there is a Hamiltonian path of length m in $G(T)$ that originates in node i and terminates in node j.

In this definition, a Hamiltonian path $i - i_1 - i_2, \ldots i_m$ of length m is an acyclic path. Since no cycles can be present then the path consists of $m+1$ *different* players.

Definition 15.2(b): The *Weighted m^{th} Generation Score* (WMGS) of player i is the number of different Hamiltonian paths of length m which originate in node i.

It is noted that the difference between these two concepts is that if k different Hamiltonian paths of length m originate in i and terminate in j, then the contribution of player j to player i is 1 under Definition 15.2(a), whereas the contribution is k under Definition 15.2(b).

Definition 15.3: The m^{th} generation adjacency matrix $A(m) = (a_{ij}(m))$ is that matrix in which $a_{ij}(m) = 1$ if there is an m-arc Hamiltonian path leading from i to j, and $a_{ij}(m) = 0$ otherwise.

The weighted adjacency matrix $A^w(m)$ is defined in the same manner except that $a_{ij}^w(m)$ is the *number* of paths from i to j.

It is noted that the m^{th} generation score for player i is the i^{th} row sum of $A(m)$. Hence,

$$a_{i.}(m) = \sum_{j=1}^{n} a_{ij}(m). \tag{15.1}$$

Similarly,

$$a_{i.}^w(m) = \sum_{j=1}^{n} a_{ij}^w(m).$$

In the following we assume that the unweighted score $a_{i.}(m)$ will be used as the measure of m^{th} order power or *strength*, rather than the weighted version, although clearly a case could be made for either of these. In order to formulate our model, it is necessary to make some assumptions:

Assumption 15.1: The score of each player is a linear combination of the m^{th} generation scores, $m = 1, 2, \ldots, n-1$. That is, each generation score contributes to the evaluation of the player.

Definition 15.4: The *weighted total rating* for player i is the sum

$$W_i = \sum_{m=1}^{n-1} \alpha_i(m) a_{i.}(m), \tag{15.2}$$

where the $\alpha_i(m)$ are the weights to be assigned to the MGS values of player i.

Assumption 15.2: The coefficients $\alpha_i(m)$ are monotonically decreasing. That is, the closer the generation, the more important is the score.

This assumption means that the scores from closer generations bear at least as much relevance as do those from more distant generations.

Assumption 15.3: Each player i should be given the opportunity to choose the best possible set of weights $\alpha_i(m)$, in the sense of rendering the rating W_i for player i as high as possible.

By defining W_i in this manner (15.2), criteria (1) and (2) are covered. Specifically, the *strength* of a player i is measured by the number of m^{th} generation wins over players j. Assumption 3 is aimed at addressing criterion (3).

With the aforementioned in mind, we consider the following model for each player i_o:

$$\max_{\{\alpha_{i_o}(m)\}_{m=1}^{n-1}} W_{i_o} = \sum_{m=1}^{n-1} \alpha_{i_o}(m) a_{i_o \cdot}(m) \qquad (15.3a)$$

subject to

$$W_i = \sum_{m=1}^{n-1} \alpha_{i_o}(m) a_{i \cdot}(m) \leq 1, i = 1, 2, \ldots, n \qquad (15.3b)$$

$$\alpha_{i_o}(1) \geq \alpha_{i_o}(2) \geq \ldots \geq \alpha_{i_o}(n-1) \geq \varepsilon > 0 \qquad (15.3c)$$

This model is solved n times, corresponding to the n players i_o. For each player, the best possible set of $\alpha_{i_o}(m)$ is chosen. The only restrictions imposed on the choice of weights are (1) that they be monotonically decreasing (15.3c), as per assumption 15.3, and (2) in the spirit of DEA, that no player rating W_i (including W_{i_o}, the one being maximized) exceed some fixed constant. For consistency, this constant is chosen as 1.

It is noted that since a different set of $\alpha(m)$ is allowable for each player, (15.3a-c) can be formulated as

$$\max \sum_{i=1}^{n} W_i = \sum_{i_o=1}^{n} \sum_{m=1}^{n-1} \alpha_{i_o}(m) a_{i_o \cdot}(m) \qquad (15.4a)$$

subject to

$$\sum_{m=1}^{n-1} \alpha_{i_o}(m) a_{i \cdot}(m) \leq 1, i_o = 1, 2, \ldots, n, i = 1, 2, \ldots, n, \qquad (15.4b)$$

$$\alpha_{i_o}(1) \geq \alpha_{i_o}(2) \geq \ldots \geq \alpha_{i_o}(n-1) \geq \varepsilon > 0, i_o = 1, 2, \ldots, n. \qquad (15.4c)$$

Model (15.4a)-(15.4c) can, therefore, be solved once to determine the set of n weights W_i, by virtue of the problem's separability.

15.2.1 Obtaining m th- Generation (Weighted m th- Generation) Scores

The weighted m^{th} generation adjacency matrix $A^w(m)$ records the number of *acyclic* paths between each pair of points in the network of players. It is noted that if one simply takes the m^{th} power A^m of the tournament matrix, the ij^{th} element gives the *total* number of m -arc paths between i and j. Unfortunately, paths with cycles are contained in this number.

An Enumeration Algorithm

In the general case, it is necessary to employ some algorithm for obtaining the acyclic paths needed in order to obtain $A^w(m)$. We suggest a simple implicit enumeration procedure as one of many possible approaches. It is noted that once $A^w(m)$ is determined, the unweighted matrix $A(m)$ is found by setting all nonzero entries equal to 1.

The Algorithm:

1. For each player i, all 1st generation paths $(i \to j)$ are obtainable from the adjacency matrix $A = (a_{ij})$. That is, if, $a_{ij} = 1$, then $i \to j$ is a 1st generation path starting at node i and ending at node j. Label node j with the set $G = \{1, 2, \ldots, n\} \setminus \{i, j\}$. So G is the set of nodes available for use in generation 2.
2. For each $(m-1)^{th}$ generation path $(i_o \to i_1 \to \ldots \to i_{m-1})$, determine the set $G = \{1, 2, \ldots, n\} \setminus \{i_o, i_1, \ldots, i_{m-1}\}$ of nodes available to be used in generation m. If G is empty or $a_{i_{m-1}j} = 0$, for all $j \varepsilon G$, go to the next $(m-1)^{th}$ generation path. If $a_{i_{m-1}j} = 1$, create the m^{th} generation path $i_o \to i_1, \to \ldots \to i_{m-1} \to j$ and revise the set G to $\bar{G} = G \setminus \{j\}$. Go to the next element of G and continue until all of its available nodes have been considered.
3. Having created all m^{th} generation paths, for each pair of players i and j, count the number of paths leading from i to j in m steps. This becomes $a_{ij}^w(m)$, the ij^{th} element of $A^w(m)$. If $m = n-1$, stop. Otherwise set $m - 1 = m$ and go to step 2.

Example:

Consider the 5-player tournament whose adjacency matrix is given by

$\underset{=}{A_1}$

	1	2	3	4	5
1	0	1	0	1	1
2	0	0	1	0	1
3	1	0	0	1	0
4	0	1	0	0	1
5	0	0	1	0	0

Figure 15-1 shows the first-generation paths. Figure 15-2 shows the second-, third-, and fourth-generation paths starting with player 2.

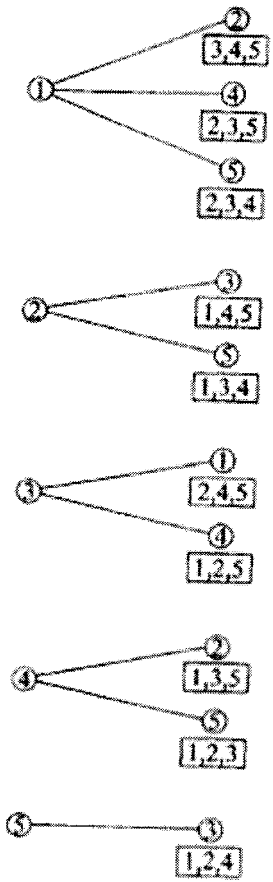

Figure 15-1. First-generation paths

Note that using the second generation node 1, only nodes 4 and 5 are allowable next (third) generation nodes. The second, third, and fourth generations out of nodes 1, 3, 4 and 5 would be determined in the same manner.

Now, to get the (2,5) entry (e.g.) of $A(3)$, we simply accumulate the number of paths arriving at node 5 at the third stage (=2), i.e., paths $2 \to 3 \to 1 \to 5$ and $2 \to 3 \to 4 \to 5$.

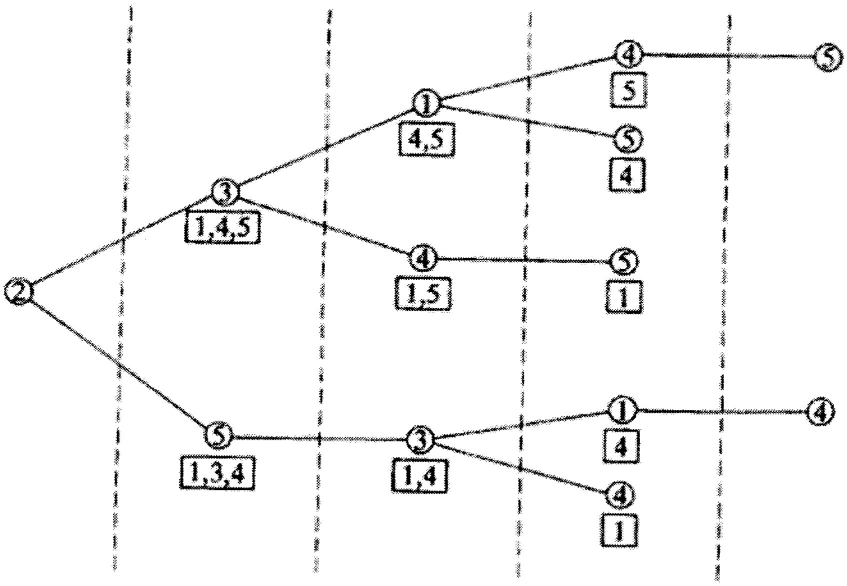

Figure 15-2. Second-, third-, and fourth-generation paths

For this tournament the m^{th} generation matrices are

$A_1^w(2)$ =

	1	2	3	4	5
1	0	1	2	0	2
2	1	0	1	1	0
3	0	2	0	1	2
4	0	0	2	0	1
5	1	0	0	1	0

$A_1^w(3)$ =

	1	2	3	4	4
1	0	0	3	2	1
2	1	0	0	2	2
3	0	1	0	0	3
4	2	0	1	0	0
5	0	2	0	1	0

$A_1^w(4)$ =

	1	2	3	4	5
1	0	1	1	1	1
2	0	0	0	1	1
3	0	0	0	0	1
4	1	1	0	0	1
5	0	1	0	0	0

15.3. WEAK RANKING OF PLAYERS

The previous section viewed each competition in the tournament as one where a direct or strict win/loss occurs. In the case where ties or draws are permitted the theory must be extended. To facilitate this extension, we introduce a broader definition of the pairwise comparison matrix to include ties or draws:

Definition 15.5: A *weak tournament* is a pairwise comparison matrix $A = (a_{ij})$ where

$$a_{ij} = \begin{cases} 1 & \text{if player } i \text{ defeats player } j, \\ 1/2 & \text{if players } i \text{ and } j \text{ tie, and} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Consider the weak tournament given by

$A_2 =$

	a	b	c	d	e	f
a	0	0	1	1/2	1/2	1
b	1	0	1/2	1	1	0
c	0	1/2	0	0	1/2	0
d	1/2	0	1	0	1/2	1/2
e	1/2	0	1/2	1/2	0	1
f	0	1	1	1/2	0	0

Definition 15.6: The *graph* G_w of a weak tournament is a graph whose set of nodes V consists of the n players and whose edges E represent the outcomes (wins, draws and losses) from the competition.

The graph G_w corresponding to the above weak tournament is shown in Figure 15-3.

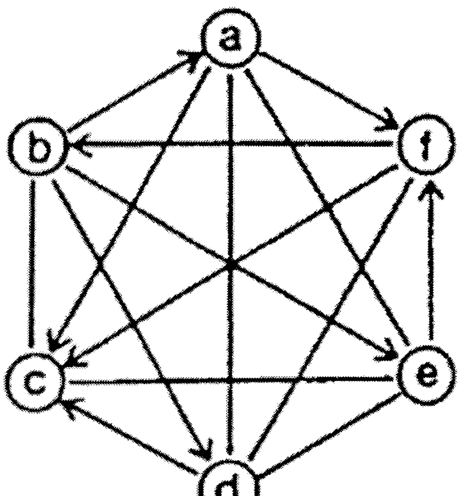

Figure 15-3. Graph G_w

Clearly G_w can consist of both directed (win or loss) and undirected (draw) edges.

The model to be presented in this section again takes account not only of immediate outcomes of player i versus player j, but outcomes in the more remote or m^{th} generation sense (i beats k who beats j). In this context, it is, therefore, necessary to look at *paths* from i to j in G_w.

Definition 15.6: The set of nodes $i_1 \geq i_2 \geq ... \geq i_k$ in G_w form a simple path of length $k-1$ if and only if $i_1, i_2, ..., i_k$ are distinct, i.e. contain no cycles.

Here, the notation $i_1 \geq i_2$ means that either i_1 defeats or ties with i_2.

As an example of a simple path, consider the aforementioned 6-player weak tournament given by A_2. The set of nodes or players *a,c,b,d,e*, for instance, constitutes a simple path of length 4 in that $a \geq c \geq b \geq d \geq e$, and no cycles exist. The path $a \geq c \geq b \geq a$ is not a simple path in that a cycle is present.

In a regular tournament (one containing no ties), the graph G contains only directed arcs, in that all preferences are strict, that is, each match produces a win. If a simple path of length m exists between two players i and j, player i is said to have an m^{th} *generation* or m^{th} order win over j. In a weak tournament, however, a simple path may contain both directed and undirected arcs. The concept of a *generation* in the pure directed graph sense is then no longer sufficient to characterize the nature of this path. The number of undirected arcs present in the path must also be accounted for. The following definition addresses this phenomenon.

Definition 15.7: A simple path in a graph G_w is said to be of type (m,k) if and only if it is of length m and the number of undirected edges in it is k, where $k \leq m$.

In order to evaluate the overall rating or rank position of a player, we wish to measure the numbers of wins of various types and orders. For this purpose, the m^{th} generation, k^{th} order scores will be utilized.

Definition 15.8: The m^{th} generation k^{th} order score (MKS) of player i is the number of players j for which there is a simple path of type (m,k) in G_w that originates in node i and terminates in node j.

In A_2 above, the 2nd generation 1st order score $(m=2, k=1)$ for player a is 5 since a has a second order win over each of b,c,d,e and f, and in each case there is a path of length 2 with exactly one pair of players in the three that tie in their match. For example, player a defeats player c who ties with player b. Thus a has a second generation first order win over b. A similar result is true for player a versus c,d,e and f.

Definition 15.9: The weighted m^{th} generation k^{th} order score (WMKS) of player i is the number of different simple paths of type (m,k) which originate in node i. It is noted that since a number of different paths can connect two players i and j, WMKS $\geq$ MKS.

The Model

Let $a_i^w(m,k)$ denote the number of distinct simple paths emanating from node i (i.e., $a_i^w(m,k)$ is the WMKS of i). Let $a_i(m,k)$ denote the MKS of i. In the next section an algorithm is given for generating the WMKS and MKS scores.

In the model to be presented herein we wish to assign each player i a score or rating which reflects his/her composite wins in a m^{th} generation sense. As well, we adopt the DEA framework for incorporating these multiple generations. In scoring a player, therefore, we make the following assumptions:

Assumption 15.4: The composite score or rating of each player is a linear combination of the WMKS. That is, each generation-order pair contributes its score to the evaluation of the player, i.e.

$$W_i = \sum_{m=1}^{n-1} \sum_{k=0}^{m} \alpha_i(m,k) a_i^w(m,k), \tag{15.5}$$

where $\alpha_i(m,k)$ is the weight to be assigned to the WMKS value of player i. The model presented below provides a mechanism for generating these weights.

It is clear that some dominance relations must be imposed on the weights $\alpha_i(m,k)$. First, it is natural to assume that within a given generation-value m, the smaller the order k the larger the weight $\alpha_i(m,k)$. Since a tie is usually considered as half a win, a reasonable requirement may be to constrain the ratio of two consecutive weights by the scores ratio of their corresponding paths, that is

$$\frac{\alpha_i(m,k)}{\alpha_i(m,k+1)} \geq \frac{m-k/2}{m-(k+1)/2}. \tag{15.6}$$

For example, if $m = 3$ (3rd generation scores), then the weight of a 1st order score (one tied preference) should be at least as large as (3-.5)/(3-1)=1.25 times the weight of a 2nd order score, where among the three relations in each of the corresponding paths, two are ties. Note that for a given k, the right hand side of (15.6) is monotonic decreasing in m, which means that as the generation gets higher the information in the corresponding scores becomes fuzzier, and, therefore, the relation between two consecutive weights is less discriminant. On the other hand, this lower bound increases, for a fixed m, as k increases. This property reflects the marginal increasing value of a strict preference when these types of preferences become scarce.

Second, the rationale for the decreasing monotonicity, for a fixed k, of the right hand side of (15.6), coupled with the dominance relation in the strict case, imply that for any $k \le m$

$$\alpha_i(m,k) \ge \alpha_i(m+1,k). \tag{15.7}$$

The above discussion is summarized in Assumption 15.5 below:

Assumption 15.5: For any m^{th} generation k^{th} order score, the corresponding weights must satisfy the following two conditions:

$$(a)\ (m-(k+1)/2)\alpha_i(m,k) - (m-k/2)\alpha_i(m,k+1) \ge 0 \tag{15.8}$$

$$(b)\ \alpha_i(m,k) \ge \alpha_i(m+1,k).$$

Under these two assumptions, let us apply the convention that weights should be chosen such that no composite score is higher than 1. Adopting the DEA approach, the following linear programming model can be utilized to derive for each player i_o, the best set of multipliers or weights $\alpha_{i_o}(m,k)$ to apply to the $a_{i_o}^w(m,k)$.

Here, $a_{io}^w(m,k)$ can denote either the MKS or WMKS:

$$\max W_{i_o} = \sum_{m=1}^{n-1}\sum_{k=0}^{m} \alpha_{i_o}(m,k) a_{i_o}^w(m,k) \tag{15.9a}$$

subject to

$$\sum_{m=1}^{n-1}\sum_{k=0}^{m} \alpha_{i_o}(m,k) a_i^w(m,k) \le 1 \qquad i = 1,\ldots,n \tag{15.9b}$$

$$(m-(k+1)/2)\alpha_{i_o}(m,k) - (m-k/2)\alpha_{i_o}(m,k+1) \ge 0$$

$$m = 1,\ldots,n-1, k = 0,\ldots,m \tag{15.9c}$$

$$\alpha_i(m,k) - \alpha_i(m+1,k) \ge 0 \quad m = 1,\ldots,n-2, k = 0,\ldots,n \tag{15.9d}$$

$$\alpha_i(n-1,n-1) \ge \varepsilon. \tag{15.9e}$$

Example: Consider the weak preference matrix A_3.

$$A_3 = \begin{array}{c|ccccc} & a & b & c & d & e \\ \hline a & 0 & 1 & 1/2 & 1 & 0 \\ b & 0 & 0 & 1 & 1 & 0 \\ c & 1/2 & 0 & 0 & 1 & 1 \\ d & 0 & 0 & 0 & 0 & 0 \\ e & 1 & 1 & 0 & 1 & 0 \end{array}$$

The WMKS values $a_i^w(m,k)$ are given by:

Table 15-1. WMKS Values for A_3

		Alternative				
m	k	a	b	c	d	e
1	0	2	2	2	—	3
	1	1	—	1	—	—
2	0	2	2	3	—	4
	1	2	1	2	—	1
	2	—	—	—	—	—
3	0	2	2	2	—	3
	1	2	1	1	—	2
	2	—	—	—	—	—
	3	—	—	—	—	—
4	0	1	1	1	—	1
	1	1	—	—	—	1
	2	—	—	—	—	—
	3	—	—	—	—	—
	4	—	—	—	—	—

It is noted that player d had no wins, hence obtains a zero score for all m and k, therefore ranking in last place.

Notice also that no scores are recorded in this example for $a^w(2,2), a^w(3,2)$, $a^w(3,3), a^w(4,2), a^w(4,3)$, and $a^w(4,4)$. Taking this observation into consideration, we can reduce the number of constraints in (15.9c)-(15.9e) to account only for relevant weights, that is, weights that apply to nonzero scores.

For example, the constraint $1.5\alpha_{i_0}(3,2) - 2\alpha_{i_0}(3.3) \geq 0$ may be omitted.

The modified problem (15.9c)-(15.9e) was solved with the data in Table 15-1, and for various values of the parameter L whose largest feasible value is .05. Table 15-2 summarizes the *optimal* composite scores for each player, and Figure 15-4 provides a graphical representation of these results.

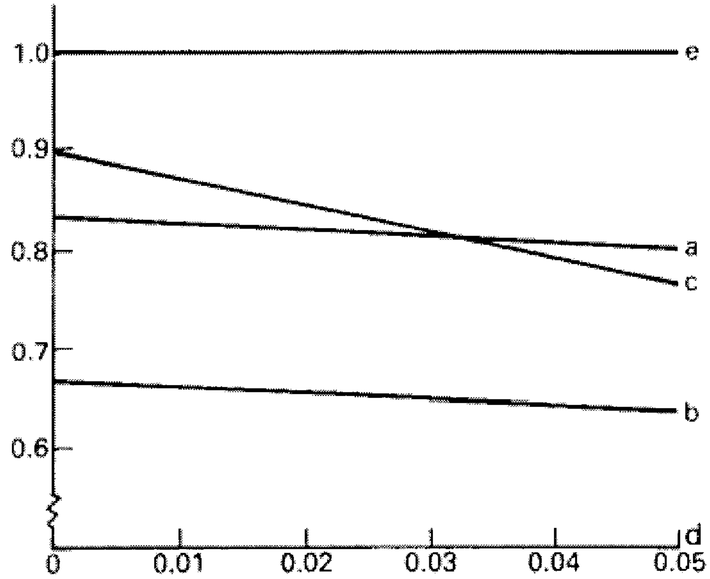

Figure 15-4. Optimal Composite Scores

Table 15-2. Optimal Composite Scores

L	0	0.01	0.02	0.03	0.04	0.05
a	0.833	0.828	0.823	0.818	0.813	0.807
b	0.667	0.656	0.644	0.633	0.622	0.608
c	0.892	0.868	0.845	0.821	0.798	0.774
d	0	0	0	0	0	0
e	1	1	1	1	1	1

For any alternative i, the composite score W_i is a non-increasing piece-wise linear function of L. In this example we note that for small values of $L(L \leq .032)$, c is rated higher than a, whereas for $L > .032$, a is rated higher than c. In any event, the tie between a and c is broken. Therefore, the overall ranking of the players is either $e > c > a > b > d$ or $e > a > c > b > d$, depending on whether higher order scores are accorded a low versus high value.

15.3.1 Obtaining WMKS

The WMKS matrix $A^w(m,k) = (a_{ij}^w(m,k))$, comprising the data of the previous problem, records the number of acyclic paths between each pair of points in the network of players. Once this matrix is determined, the WMKS value for each player i is obtained by summing the elements of the i^{th} row of the matrix. To avoid cycles, and to distinguish m-arc paths with various numbers of ties, an algorithm is required. While various approaches are possible, we suggest a simple enumeration procedure. It is noted that once $A^w(m,k)$ is determined, the unweighted matrix $A(m,k)$ is found by setting all nonzero entries to 1.

In the algorithm to follow, it is necessary to keep track of both the length m of the path and the number of ties. The latter will be recorded using flag variables $F_{i_o i_1,\dots i_m}$ = # tied players in path $i_0 \to \cdots \to i_m$.

An Enumeration Algorithm

(i) Set $F_i = 0$ for all players i.

(ii) For each player i, all 1st generation paths $(i \to j)$ are obtainable from the adjacency matrix $A = (a_{ij})$. That is, if $a_{ij} = 1$ or $1/2$ then $i \to j$ is a 1st generation path starting at node i and ending at node j. Label node j with the set $G = \{1,2,\dots,n\} \setminus \{i,j\}$. So G is the set of nodes available for use in generation 2. If $a_{ij} = 1/2$ set $F_{ij} = F_i + 1$. Otherwise $F_{ij} = F_i$.

(iii) For each $(m-1)^{th}$ generation path $(i_o \to i_1 \to \dots \to i_{m-1})$, determine the set $G = \{1,2,\dots,n\} \setminus \{i_o, i_1, \dots, i_{m-1}\}$ of nodes available to be used in generation m. If G is empty or $a_{i_{m-1}j} = 0$ for all $j \varepsilon G$, go to the next $(m-1)^{th}$ generation path. If $a_{i_{m-1}j} = 1$, create the m^{th} generation path $i_o \to i_1, \to \dots \to i_{m-1} \to j$ and revise the set G to $G = G \setminus \{j\}$. Set $F_{i_o i_1 \cdots i_{m-1} j} = F_{i_o i_1 \cdots i_{m-1}}$. If $a_{i_{m-1}j} = 1/2$ set $F_{i_o i_1 \cdots i_{m-1} j} = F_{i_o i_1 \cdots i_{m-1}} + 1$. Go to the next element of G, and continue until all of its available nodes have been considered.

(iv) Having created all m^{th} generation paths for each pair of players i and j, count the number of paths leading from i to j in m steps, and having k tied players. This becomes $a_{ij}^w(m,k)$, the ij^{th} element of $A^w(m,k)$.

If $m = n-1$, stop. Otherwise set $m-1 = m$ and go to step (iii).

Example: Consider the 5-player tournament whose adjacency matrix is given by

$A_4 =$

	1	2	3	4	5
1	0	1/2	0	1	1
2	1/2	0	1	1/2	1
3	1	0	0	1	0
4	0	1/2	0	0	1/2
5	0	0	1	1/2	0.

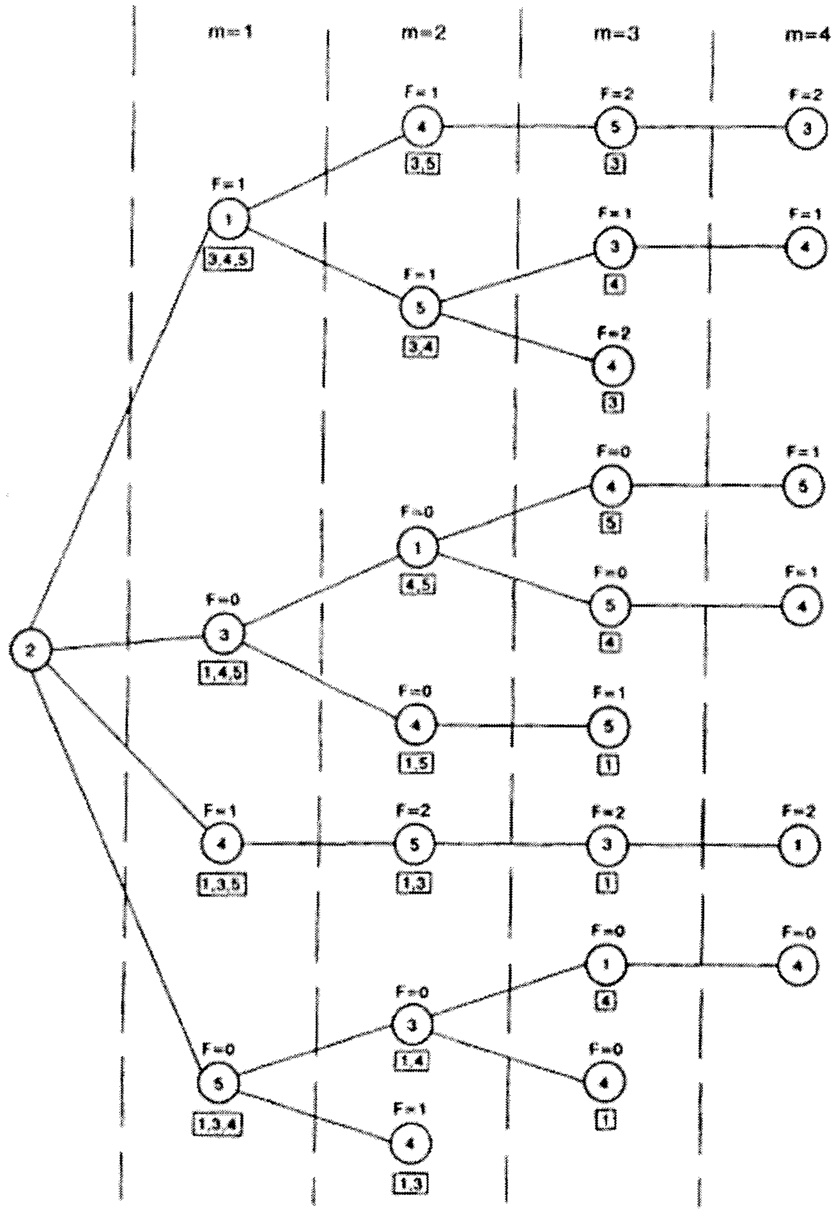

Figure 15-5. m^{th} Generation, k^{th} Order Paths for Player 2

Figure 15-5 illustrates how the algorithm is applied to obtain m^{th} generation, k^{th} order paths for a given player (player 2). For example, there is a 3rd generation, 0^{th} order path from player 2 to player 5 $(2 \to 3 \to 1 \to 5)$. There is a 3rd generation, 2nd order path from 2 to 5 as well $(2 \to 1 \to 4 \to 5)$.

A sample m^{th} generation, k^{th} order matrix is:

$A_4^w(m,k) = A_4^w(2,1) =$

	1	2	3	4	5
1	0	1	1	1	2
2	0	0	0	2	1
3	0	2	0	0	1
4	0	0	2	0	1
5	0	0	0	0	0

The corresponding WMKS values $a_i^w(2,1)$ (row sums) are:

$a_1^w\ (2,1) = 5$
$a_2^w\ (2,1) = 3$
$a_3^w\ (2,1) = 3$
$a_4^w\ (2,1) = 3$
$a_5^w\ (2,1) = 0$

15.4. MULTIPLE TOURNAMENTS

In the case where data from multiple tournaments are available the problem of rating players becomes more complex. Not only is there the fact that not all players compete in all tournaments, but there is also an issue of tournament difficulty. In the present subsection we will consider only the case of complete tournaments, i.e., where all players compete in each tournament. Tournament difficulty (or importance), which we do wish to look at here, can be viewed from at least two perspectives. First, tournament difficulty may be player dependent–competing against better players makes a tournament more difficult than would be true of one with lesser competitors. Second, the importance of a tournament may in some cases be viewed from a prestige standpoint. Wins in an internationally recognized tournament may be seen as being more important than those in lower profile matches. Thus, the issue of difficulty or importance can involve several considerations. While we do not propose to delve into the matter of how relative importance should be decided, we do wish to look at how one should evaluate players when the relative importance has been expressed.

Consider then the case of T tournaments which can be arranged in order of difficulty. Without loss of generality we assume

$$T_1 > T_2 > \cdots > T_t > \cdots T_T,$$

that is tournament T_1 is the most difficult followed by $T_2, \ldots,$ and so on. This being the case, we assume that a player is awarded more credit or weight for a win over a given player in a more difficult tournament T_{t_1} than in a less difficult tournament T_{t_2} (i.e., where $T_{t_1} > T_{t_2}$.) In that regard, let w_t be a variable (whose value is to be determined) that expresses the *level of difficulty* of tournament T_t.

In the case of multiple tournaments, several approaches can be taken to arrive at a ranking of the players. Here, our approach advocates finding a set of m th generation multipliers $\{\alpha^t(m)\}_m$ for each tournament $t = 1,\dots,T$ together with a set of tournament difficulty multipliers $\{w_t\}_t$ through the following multi tournament generalization of problem (15.3a)-(15.3c):

$$W_{i_o} = \max_{\{\alpha^t(m)\},\{w_t\}} \sum_{t=1}^{T}\sum_{m=1}^{M} w_t\alpha^t(m)a^t_{i_o}(m)$$

subject to

$$\sum_{t=1}^{T}\sum_{m=1}^{M} w_t\alpha^t(m)a^t_i(m) \le 1, i = 1,\dots,n$$

$$\sum_{m=1}^{M} \alpha^t(m)a^t_j(m) \le 1, i = 1,\dots,n; t = 1,\dots,T \tag{15.10}$$

$$\alpha^t(m) - \alpha^t(m+1) \ge \varepsilon, m = 1,\dots,M-1; t = 1,\dots,T$$

$$\alpha^t(M) \ge \varepsilon, t = 1,\dots,T$$

$$w_t - w_{t+1} \ge \varepsilon, t = 1,\dots,T-1$$

$$w_T \ge \varepsilon$$

Here, we again provide each player with the opportunity to choose not only the most favorable weights $\alpha^t(m)$ on the m generations, but also to weight the importance of the tournaments (while respecting the constraints) in a manner that makes his/her rating W_{i_o} as high as possible.

Problem (15.10), unlike problem (15.3a-c), is nonlinear in the presence of products of variables, i.e., $w_t\alpha^t(m)$. Generally, such nonlinearities would render the problem very difficult to solve, but in the present case an equivalent linear formulation is at hand.

A Linear Representation

An equivalent linear representation of problem (PT) can be accomplished through a simple change of variables. Specifically, define

$$\beta_{tm} = w_t\alpha^t(m), \tag{15.11}$$

and replace the second, third, and fourth constraints by the equivalent constraints

$$w_t \sum_{m=1}^{M} \alpha^t(m)a^t_i(m) \le w_t, i = 1,\dots,m, t = 1,\dots,T$$

$$w_t\alpha^t(m) - w_t\alpha^t(m+1) \geq \varepsilon w_t, m = 1,\dots,M-1, t = 1,\dots,T$$

$$w_t\alpha^t(M) \geq \varepsilon w_t, t = 1,\dots,T$$

Now, rewrite problem (15.10) in the form

$$R_{i_o} = \max_{\{\beta_{tm}\}} \sum_{t=1}^{T}\sum_{m=1}^{M} \beta_{tm} a_{i_o.}^t(m)$$

subject to

$$\sum_{t=1}^{T}\sum_{m=1}^{M} \beta_{tm} a_{i.}^t(m) \leq 1, i = 1,\dots,n$$

$$\sum_{m=1}^{M} \beta_{tm} a_{i.}^t(m) - w_t \leq 0, i = 1,\dots,n, t = 1,\dots,T \tag{15.12}$$

$$\beta_{tm} - \beta_{tm+1} - \varepsilon w_t \geq 0, m = 1,\dots,M-1, t = 1,\dots,T$$

$$\beta_{tM} - \varepsilon w_t \geq 0, t = 1,\dots,T$$

$$w_t - w_{t+1} \geq \varepsilon, t = 1,\dots,T-1$$

$$w_T \geq \varepsilon$$

Since all w_t are strictly positive, then an optimal solution to the LP problem (15.12) immediately yields an optimum to (15.10) in the usual way.

Common Weights Across Tournaments

In the above model a different set of multipliers $\alpha^t(m)$ arises for each tournament T_t. If it is desired to obtain a single (common) set $\alpha(m)$ that applies to all tournaments, then one would need to solve the n quadratic problems (15.10) where we replace $\alpha^t(m)$ by $\alpha(m)$. Unfortunately, in this case, the linearization procedure presented above doesn't work.

15.5. PARTIAL TOURNAMENTS

In the previous section it was assumed that each of n players participated in the same set of T tournaments. Here we examine a generalization of this idea. Assume each player $i\varepsilon\{1,\dots,n\}$ competes in some subset $K_i \subseteq \{T_1, T_2, \dots, T_T\}$ of the T tournaments. Further, it is assumed that any tournament where i competes is *complete* in the sense that each player in that tournament has exactly one match with each of the other players in the tournament. So, each of these tournaments is *round robin* in the usual sense. Finally, we assume for simplicity that the number of generations M that is used to capture player strength is common across all tournaments. As will be pointed out later, this assumption can be removed without changing the general approach.

The problem with attempting to model player performance in the partial tournament setting using (15.10) is that a player i_o who competes in only a small subset K_{i_o} of tournaments will tend to be dominated by a player i_1, where K_{i_1} is a much larger set; i.e., $W_{i_o} < W_{i_1}$, simply because of the numbers of matches played in the two cases. Hence, this formulation fails to account for the differential numbers of matches K_i played. To accomplish this we propose the following generalization of (15.10)

$$W_{i_o} = \max_{\{\alpha^t(m)\},\{w_t\}} \frac{\sum_{t\varepsilon K_{i_o}} \sum_{m=1}^{M} w_t \alpha^t(m) a^t_{i_o\cdot}(m)}{\sum_{t\varepsilon K_{i_o}} w_t}$$

subject to

$$\frac{\sum_{t\varepsilon K_i} \sum_{m=1}^{M} w_t \alpha^t(m) a^t_{i_o}(m)}{\sum_{t\varepsilon K_i} w_t} \le 1, i = 1,\dots,n$$

$$\sum_{m=1}^{M} \alpha^t(m) a^t_{i}(m) \le 1, i = 1,\dots,n, t\varepsilon K_i$$

$$\alpha^t(m) - \alpha^t(m+1) \ge \varepsilon, m = 1,\dots,M-1, t = 1,\dots,T \quad (15.13)$$

$$\alpha^t(m) \ge \varepsilon, t = 1,\dots,T$$

$$w_t - w_{t+1} \ge \varepsilon, t = 1,\dots,T-1$$

$$w_T \ge \varepsilon$$

Problem (15.13), a fractional linear problem, therefore accounts for tournament participation by way of normalization. In this manner, players can be properly compared regardless of the number of tournaments in which each is involved.

Following the earlier change of variables $\beta_{tm} = \alpha^t(m) w_t$, and letting $v_t = \tau_{i_o} w_t$ and $\mu_{tm} = \tau_{i_o} \beta_{tm}$, problem (15.13) is equivalent to the linear programming problem

$$W_{i_o} = \max \sum_{t\varepsilon K_{i_o}} \sum_{m=1}^{M} \mu_{tm} a^t_{i_o\cdot}(m)$$

subject to

$$\sum_{t\varepsilon K_{i_o}} v_t = 1$$

$$\sum_{t\varepsilon K_i} \sum_{m=1}^{M} \mu_{tm} a^t_{i\cdot}(m) - \sum_{t\varepsilon K_i} v_t \le 0, i = 1,\dots,n$$

$$\sum_{m=1}^{M} \mu_{tm} a_{i.}^{t}(m) - v_t \leq 0, i = 1,\dots,n, t\varepsilon K_i \tag{15.14}$$

$$\mu_{tm} - \mu_{tm+1} \geq \varepsilon v_t, m = 1,\dots,M-1, t = 1,\dots,T$$

$$\mu_{tM} \geq \varepsilon v_t, t = 1,\dots,T$$

$$v_t - v_{t+1} \geq \varepsilon \tau_{i_o}, t = 1,\dots,T-1$$

$$v_T \geq \varepsilon \tau_{i_o}$$

In solving problem (15.13) we may impose a restriction of the form $\sum_{t\varepsilon K_i} w_t \leq \ominus$ for some chosen scaler $\ominus$. Since ε is an infinitesimal, we may with no loss of generality choose $\ominus = 1$. The maximum value of W_{i_o} for $\ominus > 1$ is only greater than W_{i_o} at $\ominus = 1$ by an amount of the order of ε. That is,

$$W_{i_o}(\ominus > 1) = W_{i_o}(\ominus) + 0(\varepsilon).$$

Thus, from a practical point of view, problem (15.13) is equivalent to a problem with a constraint $\sum_{t\varepsilon k_i} w_t \leq 1$ added.

Thus, we may augment problem (15.14) by the additional restriction

$$\tau_{i_o} \geq 1 \tag{15.15}$$

Again, given a solution (τ_o^*, μ^*, v^*) to (15.15), then $w_t^* = \frac{v_t^*}{\tau_o^*}$ and $\alpha^{t*}(m) = \frac{\mu_{tm}}{v_t^*}$ is a solution to (15.13). It is noted that at the optimum $\tau_o^* = \overline{\tau}_{i_o}$.

15.6. CONCLUSIONS

This chapter has examined the application of the DEA model to the problem of ranking players in a tournament. Tournament theory has a long history with many different methodologies having been applied to the player ranking issue. These previous methods have generally been rating-based or have involved maximum-likelihood concepts. The advantage of the DEA approach is the maximum fairness principle that it embodies. It also offers the opportunity to directly address extensions to the original tournament structures, namely, those pertaining to multiple matches and partial participation in those matches.

REFERENCES

1. Ali, I., Wade D. Cook and M. Kress (1986), On the minimum violations ranking of a tournament, *Management Science* 32(6), 660- 672.

2. Charnes, A., W.W. Cooper and E. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2(6), 429-444.
3. Cook, W.D., I. Golan, A. Kazakov and M. Kress (1988c), A case study of a noncompensatory approach to ranking transportation projects, *Journal of the Operational Research Society* 39(11), 901-910.
4. Cook, W.D., I. Goland and M. Kress (1988a), Heuristics for ranking players in a round robin tournament, *Computers and Operations Research* 15(2), 135-144.
5. Cook, W.D. and M. Kress (1988b), Partial and multiple match tournaments, *Mathematical Social Sciences* 15, 303-306.
6. Cook, W.D. and M. Kress (1990), An m^{th} generation model for weak ranking of players in a tournament, *Journal of the Operational Research Society* 41(2), 1111-1119.
7. Cook, W.D. and M. Kress (1992), Ranking players in a tournament: an m^{th} generation approach, *Systems and Management Science by Extremal Methods,* Kluwer Academic Publishers, Boston.
8. Goddard, S.T. (1983), Ranking tournaments and group decision making, *Management Science* 29(12), 1384-1392.
9. Kendall, M. (1962), Rank Correlation Methods, Hafner, New York.
10. Moon, J.W. (1968), Topics on Tournaments, Holt, Rinehard and Winston, New York.
11. Roy, B. (1968), Classement et choix en presence de points de vue multiples (la methode electre), *Revue Francaise Automatique, Informatique et Recherche Operationnelle,* 8, 57-75.
12. Wei, T.H., The algebraic foundations of ranking theory, PhD Thesis, Cambridge University.

This chapter is based upon (i) Cook, W.D. and M. Kress (1990), An m^{th} Generation Model for Weak Ranking of Players in a Tournament, Journal of the Operational Research Society, Vol. 41, No. 12, 1111-1119, with permission from Palgrave, (ii) Cook, W.D. and M. Kress (1992), Rating Players in a Tournaments; An m^{th} Generation Approach, in Systems and Management Science by Extremal Methods, Kluwer Academic Publishers, Boston, with permission from Kluwer Academic Publishers, and (iii) Cook, W.D., J. Doyle, R. Green and M. Kress (1996), Ranking Players in Multiple Tournaments, Computers and Operations Research, Vol. 24, 392-397 with permission from Elsiever Publishing.

Chapter 16

CONTEXT-DEPENDENT DEA

Models and Extension

16.1. INTRODUCTION

DEA identifies efficient DMUs from a given set of DMUs. It is well known that adding or deleting an inefficient DMU or a set of inefficient DMUs does not alter the efficiencies of the existing DMUs and the efficient frontier. The inefficiency scores change only if the efficient frontier is altered. i.e., the performance of DMUs depends only on the identified efficient frontier. In contrast, researchers of the consumer choice theory point out that consumer choice is often influenced by the context. e.g., a circle appears large when surrounded by small circles and small when surrounded by larger ones. Similarly, a product may appear attractive against a background of less attractive alternatives and unattractive when compared to more attractive alternatives (Simonson and Tversky, 1992).

Considering this influence within the framework of DEA, one could ask "what is the relative attractiveness of a particular DMU when compared to others?" As in Tversky and Simonson (1993), one agrees that the relative attractiveness of DMU_x compared to DMU_y, say DMU_z (or a group of DMUs). Relative attractiveness depends on the evaluation context constructed from alternative options (or DMUs).

In fact, a set of DMUs can be divided into different levels of efficient frontiers. If we remove the (original) efficient frontier, then the remaining (inefficient) DMUs will form a new second-level efficient frontier. If we remove this new second-level efficient frontier, a third-level efficient frontier is formed, and so on, until no DMU is left. Each such efficient frontier provides an evaluation context for measuring the relative attractiveness. e.g., the second-level efficient frontier serves as the evaluation context for measuring the relative attractiveness of the DMUs located on the first-level (original) efficient frontier. On the other hand, we can measure the performance of DMUs on the third-level efficient frontier with respect to the first or second level efficient frontier.

In this way, we obtain a context-dependent DEA where the relative attractiveness is obtained when DMUs having worse performance are chosen as the evaluation context, and the relative progress is obtained when DMUs having better performance are chosen as the evaluation context. The presence or absence (or the shape) of the evaluation context (efficient frontier) affects the relative attractiveness or progress of DMUs on a different level of efficient frontier. When DMUs in a specific level are viewed as having equal performance, the attractiveness measure or the progress measure allows us to differentiate the "equal performance" based upon the same specific evaluation context (or third option).

Note that different input/output measures play different roles in the evaluation of a DMU's performance. Customers may make trade-offs among different measures of a product. For example, suppose we want to buy a dot-matrix printer and we may, given the price, make tradeoffs amongst the speed, print quality, and input buffer (memory) which are some of the most important features that distinguish 24-pin dot-matrix printers. We may not consider the printer memory feature to be very vital, because dot-matrix printers only use memory as a buffer space to download fonts. Thus, we give more consideration to speed and print quality. Perhaps, the printer is simply used to print long program codes or data-base listings, so that speed outweighs print quality.

Therefore, in measuring the relative attractiveness and progress, incorporation of value judgment is also very important. The current chapter uses the result of Zhu (2002) to develop a context-dependent DEA with value judgment. The method is applied to measure the relative attractiveness of a set of printers that is studied by Doyle and Green (1991). The application demonstrates that the context-dependent DEA helps practitioners to produce finer evaluation of efficiency in practical problems.

The rest of the chapter is organized as follows. The next section presents the context-dependent DEA. We then incorporate the value judgment into the context-dependent DEA. The method is applied to a set of 32 printers. Conclusions are provided in the last section.

16.2. CONTEXT-DEPENDENT DEA

Our model formulation below uses a vector notion for inputs and outputs where DMU_j (j = 1, 2, ..., n) produces a vector of outputs $y_j = (y_{1j}, \ldots, y_{sj})$ by using a vector of inputs $x_j = (x_{1j}, \ldots, x_{mj})$.

Let $\mathbf{J}^1 = \{DMU_j, j = 1, \ldots, n\}$ be the set of all n DMUs. We interactively define $\mathbf{J}^{l+1} = \mathbf{J}^l - \mathbf{E}^l$ where $\mathbf{E}^l = \{DMU_k \in \mathbf{J}^l \mid \phi^*(l,k) = 1\}$, and $\phi^*(l,k)$ is the optimal value to the following linear programming problem:

$$\phi^*(l,k) = \max_{\lambda_j, \phi(l,k)} \phi(l,k)$$

$$\text{subject to}$$

$$\sum_{j \in \mathrm{F}(\mathbf{J}^l)} \lambda_j y_j \geq \phi(l,k) y_k; \tag{16.1}$$

$$\sum_{j \in \mathrm{F}(\mathbf{J}^l)} \lambda_j x_j \leq x_k;$$

$$\lambda_j \geq 0 \quad j \in \mathrm{F}(\mathbf{J}^l)$$

where (x_k, y_k) represents the input and output vector of DMU_k, and $j \in \mathrm{F}(\mathbf{J}^l)$ means $DMU_j \in \mathbf{J}^l$, i.e., $\mathrm{F}(\cdot)$ represents the correspondence from a DMU set to the corresponding subscript index set.

When l=1, model (1) becomes the original output-oriented CCR model and DMUs in set $\mathbf{E}^1$ define the first-level efficient frontier. When l = 2, model (1) gives the second-level efficient frontier after the exclusion of the first-level efficient DMUs. And so on. In this manner, we identify several levels of efficient frontiers. We call $\mathbf{E}^l$ the lth-level efficient frontier. The following algorithm accomplishes the identification of these efficient frontiers by model (1).

- ***Step 1***: Set l = 1. Evaluate the entire set of DMUs, $\mathbf{J}^1$, by model (16.1) to obtain the first-level efficient DMUs, set $\mathbf{E}^1$ (the first-level efficient frontier).
- ***Step 2***: Exclude the efficient DMUs from future DEA runs. $\mathbf{J}^{l+1} = \mathbf{J}^l - \mathbf{E}^l$. (If $\mathbf{J}^{l+1} = \varnothing$ then stop.)
- ***Step 3***: Evaluate the new subset of "inefficient" DMUs, $\mathbf{J}^{l+1}$, by model (1) to obtain a new set of efficient DMUs $\mathbf{E}^{l+1}$ (the new efficient frontier).
- ***Step 4***: Let $l = l + 1$. Go to step 2.
- ***Stopping rule***: $\mathbf{J}^{l+1} = \varnothing$, the algorithm stops.

There exists an input-oriented version of model (16.1). However, the input-oriented version of model (16.1) yields the same stratification of the whole set of DMUs. Figure 16.1 plots the three levels of efficient frontiers of 10 DMUs with two outputs and one single input of one (see Table 16.1).

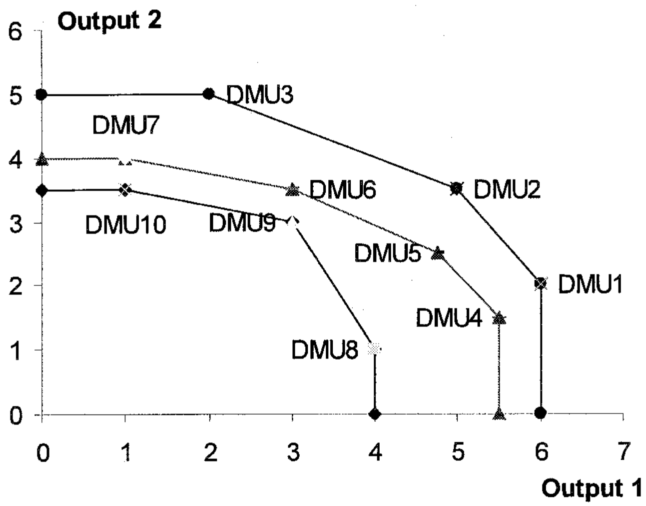

Figure 16-1. Efficient Frontiers

Table 16-1. Sample DMUs

DMU	1	2	3	4	5	6	7	8	9	10
Output 1	6	5	2	5.5	4.75	3	1	4	3	1
Output 2	2	3.5	5	1.5	2.5	3.5	4	1	3	3.5

Now, based upon these evaluation contexts $\mathbf{E}^l$ (l = 1, ..., L), we can obtain the relative attractiveness measure by the following context-dependent DEA.

$$\Omega_q^*(d) = \max_{\lambda_j, \Omega_q(d)} \Omega_q(d) \qquad d = 1,\dots, L - l_o$$

subject to

$$\sum_{j \in \mathrm{F}(\mathbf{E}^{l_o+d})} \lambda_j y_j \ge \Omega_q(d) y_q; \qquad (16.2)$$

$$\sum_{j \in \mathrm{F}(\mathbf{E}^{l_o+d})} \lambda_j x_j \le x_q;$$

$$\lambda_j \ge 0 \quad j \in \mathrm{F}(\mathbf{E}^{l_o+d}).$$

where $DMU_q = (x_q, y_q)$ is from a specific level $\mathbf{E}^{l_o}$, $l_o \in \{1, ..., L\text{-}1\}$. We have (i) $\Omega_q^*(d) < 1$ for each $d = 1,\dots, L - l_o$, and (ii) $\Omega_q^*(d+1) < \Omega_q^*(d)$.

Definition 16.1: $\mathrm{A}_q^*(d) \equiv \dfrac{1}{\Omega_q^*(d)}$ is called the (output-oriented) *d-degree* attractiveness of DMU_q from a specific level $\mathbf{E}^{l_o}$.

Suppose, e.g., each DMU in the first-level efficient frontier represents an option, or product. Customers usually compare a specific DMU in $\mathbf{E}^{l_o}$ with other alternatives that are currently in the same level as well as with relevant alternatives that serve as evaluation contexts. The relevant alternatives are those DMUs, say, in the second or third level efficient frontier, etc.. Given the alternatives (evaluation contexts), model (16.2) enables us to select the best option – the most attractive one.

In model (16.2), each efficient frontier of $\mathbf{E}^{l_o+d}$ represents an evaluation context for measuring the relative attractiveness of DMUs in $\mathbf{E}^{l_o}$. Note that $\mathrm{A}_q^*(d)$ is the reciprocal of the optimal value to (16.2), therefore $\mathrm{A}_q^*(d) > 1$. The larger the value of $\mathrm{A}_q^*(d)$, the more attractive the DMU_q is, because this DMU_q makes itself more distinctive from the evaluation context $\mathbf{E}^{l_o+d}$. We are able to rank the DMUs in $\mathbf{E}^{l_o}$ based upon their attractiveness scores and identify the best one.

To obtain the progress measure for a specific $DMU_q \in \mathbf{E}^{l_o}$, $l_o \in \{2, ..., L\}$, we use the following context-dependent DEA.

$$
\begin{aligned}
& P_q^*(g) = \max_{\lambda_j, P_q(g)} P_q(g) \qquad g = 1,\dots, l_o - 1 \\
& \text{subject to} \\
& \sum_{j \in \mathrm{F}(E^{l_o-g})} \lambda_j y_j \geq P_q(g) y_q; \\
& \sum_{j \in \mathrm{F}(E^{l_o-g})} \lambda_j x_j \leq x_q; \\
& \lambda_j \geq 0 \quad j \in \mathrm{F}(E^{l_o-g}).
\end{aligned}
\tag{16.3}
$$

We have (i) $P_q^*(g) > 1$ for each $g = 1, \dots, l_o\text{-}1$, and (ii) $P_q^*(g+1) > P_q^*(g)$.

Definition 16.2: The optimal value to (16.3), i.e., $P_q^*(g)$, is called the (output-oriented) *g-degree* progress of DMU_q from a specific level $\mathbf{E}^{l_o}$.

Each efficient frontier, $\mathbf{E}^{l_o-g}$, contains a possible target for a specific DMU in $\mathbf{E}^{l_o}$ to improve its performance. The progress here is a level-by-level improvement. For a larger $P_q^*(g)$, more progress is expected for DMU_q. Thus, a smaller value of $P_q^*(g)$ is preferred.

16.3. CONTEXT-DEPENDENT DEA WITH VALUE JUDGMENT

In the previous section, both attractiveness and progress are measured radially with respect to different *levels* of efficient frontiers. The measurement does not require *a priori* information on the importance of the attributes (input/output) that feature the performance of DMUs. However different attributes play different roles in the evaluation of a DMU's overall performance. Therefore, we introduce value judgment into the context-dependent DEA.

Incorporating Value Judgment into Attractiveness Measure

In order to incorporate such *a priori* information into our measures of attractiveness and progress, we first specify a set of weights related to the s outputs, u_r (r = 1, ..., s) such that $\sum u_r = 1$. Based upon Zhu (1996), we develop the following linear programming problem for $DMU_q = (x_q, y_q)$ $= (x_{1q},\ldots,x_{mq}, y_{1q},\ldots,y_{sq})$ in $\mathbf{E}^{l_o}$, $l_o \in \{1, ..., \text{L-1}\}$:

$$\Phi_q^*(d) = \max_{\lambda_j, \Phi_q^r(d)} \sum_{r=1}^{s} u_r \Phi_q^r(d) \qquad d = 1,\ldots,L-l_o$$

$$\begin{aligned} &\text{subject to} \\ &\sum_{j\in \mathrm{F}(\mathbf{E}^{l_o+d})} \lambda_j y_{rj} \geq \Phi_q^r(d) y_{rq} \qquad r = 1,\ldots,s; \\ &\sum_{j\in \mathrm{F}(\mathbf{E}^{l_o+d})} \lambda_j x_{ij} \leq x_{iq} \qquad i = 1,\ldots,m; \\ &\Phi_q^r(d) \leq 1 \qquad r = 1,\ldots,s; \\ &\lambda_j \geq 0 \quad j \in \mathrm{F}(\mathbf{E}^{l_o+d}). \end{aligned} \tag{16.4}$$

Definition 16.4: $\overline{\overline{A}}_q^*(d) \equiv \dfrac{1}{\Phi_q^*(d)}$ is called the (output-oriented) value judgment (VJ) *d-degree* attractiveness of DMU_q from a specific level $\mathbf{E}^{l_o}$.

Obviously, $\overline{\overline{A}}_q^*(d) > 1$. The larger the $\overline{\overline{A}}_q^*(d)$ is, the more attractive the DMU_q appears under the weights u_r (r = 1,..., s). We now can rank DMUs in the same level by their VJ attractiveness scores incorporated with the preferences over outputs.

If one wishes to prioritize the options (DMUs) with higher values of the $r_o th$ output, then one can increase the value of the corresponding weight u_{r_o}. These user-specified weights reflect the relative degree of desirability of the corresponding outputs. For example, if one prefers a printer with faster printing speed to one with higher print quality, then one may specify a larger weight for the speed (output). The constraints of $\Phi_q^r(d) \le 1$ (r = 1, ..., s) ensure that in an attempt to make itself as distinctive as possible, DMU_q is not allowed to decrease some of its outputs to achieve higher levels of other preferred outputs.

Consider DMUs, 1, 2 and 3 in Table 16.1 and select the second-level efficient frontier as the evaluation background, i.e., we consider the VJ first degree attractiveness.

Case I: If let $u_1 = u_2 = 0.5$, i.e., the preference over the two outputs is equal, then we have $\overline{\overline{A}}_1^*(1) = 1.0787$, $\overline{\overline{A}}_2^*(1) = 1.2019$ and $\overline{\overline{A}}_3^*(1) = 1.1429$. Thus, DMU2 is the most attractive one;

Case II: If let $u_1 = 0.98$ and $u_2 = 0.02$, i.e., we prefer the first output, then we have $\overline{\overline{A}}_1^*(1) = 1.0949$, $\overline{\overline{A}}_2^*(1) = 1.0077$ and $\overline{\overline{A}}_3^*(1) = 1.0050$. Thus, DMU1 is the most attractive one;

Case III: If $u_1 = 0.02$ and $u_2 = 0.98$, i.e., we prefer the second output, then we have $\overline{\overline{A}}_1^*(1) = 1.0030$, $\overline{\overline{A}}_2^*(1) = 1.0081$ and $\overline{\overline{A}}_3^*(1) = 1.2595$. Thus, DMU3 is the most attractive one.

It can be seen that different weight combinations lead to different attractiveness scores.

Note that $\overline{\overline{A}}_q^*(d)$ (or $\Phi_q^*(d)$) is an overall attractiveness of DMU_q in terms of outputs while keeping the inputs at their current levels. On the other hand, each individual optimal value of $\frac{1}{\Phi_q^{r*}(d)}$, (r = 1, …, s) measures the attractiveness of DMU_q in terms of each output dimension. Note that $\overline{\overline{A}}_q^*(d)$ is not equal to $\sum_{r=1}^{s} u_r A_q^r(d)$, where $A_q^r(d) = \frac{1}{\Phi_q^{r*}(d)}$.

Definition 16.5: For $DMU_q \in \mathbf{E}^{l_o}$, $l_o \in \{2, ..., L\}$, the optimal value

$$\overline{\overline{A}}_q^{r*}(d) \equiv \frac{1}{\Phi_q^{r*}(d)}$$

is called the (output-oriented) value judgment (VJ) *d-degree* output-specific attractiveness measure.

Consider case I of VJ first degree attractiveness. When $u_1 = u_2 = 0.5$, we have (i) $A_1^1(1) = 1.1710$, $A_1^2(1) = 1$ for DMU1; (ii) $A_2^1(1) = 1.0526$, $A_2^2(1) = 1.4006$ for DMU2; and (iii) $A_3^1(1) = 1$, $A_3^2(1) = 1.3333$ for DMU3. Thus, DMU1 is the most attractive one in terms of the first output, whereas DMU2 is the most attractive one in terms of the second output.

Let $\Phi_q^r(d)y_{rq} = y_{rq} - s_q^r(d)$ (r = 1, ..., s) in (4). Since $\Phi_q^r(d) \le 1$, $s_q^r(d) \ge 0$, model (4) is equivalent to the following linear programming problem:

$$\begin{aligned} &\min_{\lambda_j, s_q^r(d)} \sum_{r=1}^{s} D_r s_q^r(d) && d = 1,\dots,L-l_o \\ s.t.\quad & y_{rq} - \sum_{j\in \mathrm{F}(\mathbf{E}^{l_o+d})} \lambda_j y_{rj} = s_q^r(d) && r = 1,\dots,s; \\ & \sum_{j\in \mathrm{F}(\mathbf{E}^{l_o+d})} \lambda_j x_{ij} \le x_{iq} && i = 1,\dots,m; \\ & s_q^r(d) \ge 0 && r = 1,\dots,s; \\ & \lambda_j \ge 0 \quad j \in \mathrm{F}(\mathbf{E}^{l_o+d}). \end{aligned} \tag{16.5}$$

where $D_r = \dfrac{u_r}{y_{rq}}$, i.e., u_r is normalized by the corresponding output quantity. $s_q^r(d)$ in (5) can be regarded as the maximum possible output reduction to a specific efficient frontier $\mathbf{E}^{l_o+d}$. Therefore, the output-specific attractiveness measure characterizes the difference between $DMU_q \in \mathbf{E}^{l_o}$ and $\mathbf{E}^{l_o+d}$ in terms of a specific output.

With the output-specific (or input-specific) attractiveness measures, one can further identify which outputs (inputs) play important roles in distinguishing a DMU's performance. On the other hand, if $\Phi_q^{r_o*}(d) = 1$, then other DMUs in $\mathbf{E}^{l_o+d}$ or their combinations can also produce the amount of the r_oth output of DMU_q, i.e., DMU_q does not exhibit better performance with respect to this specific output dimension. Therefore, DMU_q should improve its performance on the r_oth output to distinguish itself in the future.

Incorporating Value Judgment into Progress Measure

Similar to the development in the previous section, we can define the output-oriented value judgment (VJ) progress measure:

$$\overline{\overline{P}}_q^*(g) = \max_{\lambda_j, P_q^r(g)} \sum_{r=1}^{s} u_r P_q^r(g) \qquad g = 1,\ldots,l_o - 1$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{j \in F(\mathbf{E}^{l_o - g})} \lambda_j y_{rj} \geq P_q^r(g) y_{rq} && r = 1,\ldots,s; \\ & \sum_{j \in F(\mathbf{E}^{l_o - g})} \lambda_j x_{ij} \leq x_{iq} && i = 1,\ldots,m; \\ & P_q^r(g) \geq 1 && r = 1,\ldots,s; \\ & \lambda_j \geq 0 \quad j \in F(\mathbf{E}^{l_o - g}). \end{aligned} \tag{16.6}$$

Definition 16.6: The optimal value $\overline{\overline{P}}_q^*(g)$ is called the (output-oriented) value judgment (VJ) *g-degree* progress of DMU_q in a specific level $\mathbf{E}^{l_o}$.

The larger the $\overline{\overline{P}}_q^*(g)$ is, the greater the amount of progress is expected for DMU_q. Here the user-specified weights reflect the relative degree of desirability of improvement on the individual output levels.

16.4. INPUT-ORIENTED CONTEXT-DEPENDENT DEA

Here we provide the input-oriented context-dependent DEA. Consider the following linear programming problem for $DMU_q = (x_q, y_q)$ in a specific level $\mathbf{E}^{l_o}$, $l_o \in \{1, \ldots, \text{L-1}\}$:

$$H_q^*(d) = \min H_q(d) \qquad d = 1,\ldots,L - l_o$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{j \in F(\mathbf{E}^{l_o + d})} \lambda_j x_j \leq H_q(d) x_q; \\ & \sum_{j \in F(\mathbf{E}^{l_o + d})} \lambda_j y_j \geq y_q; \\ & \lambda_j \geq 0 \quad j \in F(\mathbf{E}^{l_o + d}). \end{aligned} \tag{16.7}$$

Note that dividing each side of the constraint of (A.1) by $H_q(d)$ gives:

$$\sum_{j\in F(\mathbf{E}^{l_o+d})} \tilde{\lambda}_j x_j \le x_q$$

$$\sum_{j\in F(\mathbf{E}^{l_o+d})} \tilde{\lambda}_j y_j \ge \frac{1}{H_q(d)} y_q$$

$$\tilde{\lambda}_j = \frac{\lambda_j}{H_q(d)} \ge 0 \quad j \in F(\mathbf{E}^{l_o+d})$$

Therefore (A.1) is equivalent to (2), and we have (i) $H_q^*(d) = \dfrac{1}{\Omega_q^*(d)}$ for $DMU_q \in \mathbf{E}^{l_o}$, $l_o \in \{1, ..., L\text{-}1\}$, (ii) $H_q^*(d) > 1$ for each $d = 1,\ldots,L-l_o$, and $H_q^*(d+1) > H_q^*(d)$.

Definition 16.7: $H_q^*(d)$ is called (input-oriented) *d-degree* attractiveness of DMU_q from a specific level $\mathbf{E}^{l_o}$.

The bigger the $H_q^*(d)$ is, the more attractive the DMU_q is. Model (A.1) determines the relative attractiveness score for DMU_q when outputs are fixed at their current levels. To measure the progress of $DMU_q \in \mathbf{E}^{l_o}$, $l_o \in \{2, ..., L\}$, we develop

$$\begin{aligned} G_q^*(g) &= \min G_q(g) \qquad g = 1,\ldots,l_o - 1 \\ \text{s.t.} \quad & \sum_{j\in F(\mathbf{E}^{l_o-g})} \lambda_j x_j \le G_q(\beta) x_q; \\ & \sum_{j\in F(\mathbf{E}^{l_o-g})} \lambda_j y_j \ge y_q; \\ & \lambda_j \ge 0 \quad j \in F(\mathbf{E}^{l_o-g}). \end{aligned} \tag{16.8}$$

We have (i) $G_q^*(g) = \dfrac{1}{P_q^*(g)}$ for $DMU_q \in \mathbf{E}^{l_o}$, $l_o \in \{2, ..., L\}$, (ii) $G_q^*(g) < 1$ for each $g = 1,\ldots,l_o - 1$, and (iii) $G_q^*(g+1) < G_q^*(g)$.

Definition 16.8: $M_q^*(g) \equiv \dfrac{1}{G_q^*(g)}$ is called (input-oriented) *g-degree* progress of DMU_q from a specific level $\mathbf{E}^{l_o}$.

Obviously $M_q^*(g) > 1$. For a larger $M_q^*(g)$, more progress is expected. Next, we develop the following linear programming problem for $DMU_q = (x_q, y_q) = (x_{1q},\ldots,x_{mq}, y_{1q},\ldots,y_{sq})$ in $\mathbf{E}^{l_o}$, $l_o \in \{1, ..., L\text{-}1\}$:

$$\overline{\overline{H}}_q^*(d) = \min \sum_{i=1}^{m} w_i H_q^i(d) \quad d = 1,\dots,L-l_o$$
$$\begin{aligned} s.t. \quad & \sum_{j \in F(\mathbf{E}^{l_o+d})} \lambda_j x_{ij} \le H_q^i(d) x_{iq} && i = 1,\dots,m; \\ & \sum_{j \in F(\mathbf{E}^{l_o+d})} \lambda_j y_{rj} \ge y_{rq} && r = 1,\dots,s; \\ & H_q^i(d) \ge 1 && i = 1,\dots,m; \\ & \lambda_j \ge 0 \quad j \in F(\mathbf{E}^{l_o+d}). \end{aligned}$$

where w_i (i = 1, ..., m) such that $\sum_{i=1}^{m} w_i = 1$ are user-specified weights reflecting the preference over the input improvements.

Definition 16.9: The optimal value $\overline{\overline{H}}_q^*(d)$ is called (input-oriented) value judgment (VJ) *d-degree* attractiveness of DMU_q in a specific level $\mathbf{E}^{l_o}$

To measure the (input-oriented) value judgment progress, we have

$$\overline{\overline{G}}_q^*(g) = \min \sum_{i=1}^{m} w_i G_q^i(g) \quad g = 1,\dots,l_o - 1$$
$$\begin{aligned} \text{s.t.} \quad & \sum_{j \in F(\mathbf{E}^{l_o-g})} \lambda_j x_{ij} \le G_q^i(g) x_{iq}; && i = 1,\dots,m; \\ & \sum_{j \in F(\mathbf{E}^{l_o-g})} \lambda_j y_{rj} \ge y_{rq}; && r = 1,\dots,s; \\ & G_q^i(g) \le 1 && i = 1,\dots,m; \\ & \lambda_j \ge 0 \quad j \in F(\mathbf{E}^{l_o-g}). \end{aligned}$$

Definition 16.10: The optimal value $\overline{\overline{M}}_q^*(g) \equiv \dfrac{1}{\overline{\overline{G}}_q^*(g)}$ ($\underline{\Delta}$), is the (input-oriented) value judgment (VJ) *g-degree* progress of DMU_q from a specific level $\mathbf{E}^{l_o}$.

16.5. CONTEXT-DEPENDENT DEA MODELS IN DEAFRONTIER SOFTWARE

Here, we demonstrate how these context-dependent DEA models can be solved using the DEAFrontier software. The context-dependent DEA consists of three functions: (i) Obtain levels, (ii) Calculate context-depend DEA models, and (iii) Unprotect the sheets containing the levels (see Figure 16-2).

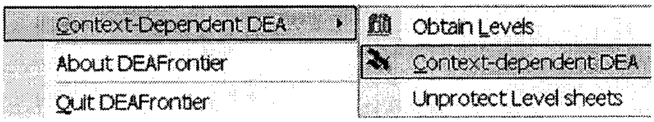

Figure 16-2. Context-dependent DEA Menu

The first function is the stratification model (16.1). It generates all the efficient frontiers – levels (Figure 16-3). This function will first delete any sheet with a name starting with "Level" and then generate a set of new sheets named as "Level*i*(*Frontier*)" where *i* indicates the level and *Frontier* represents the frontier type. For example, Level1(CRS) means the first level CRS frontier. The "level" sheets are protected for use in the context-dependent DEA. However, they can be unprotected by using the "Unprotect the sheets" menu item. *The format of these level sheets must not be modified. Otherwise, the context-dependent DEA will not run properly and accurately.*

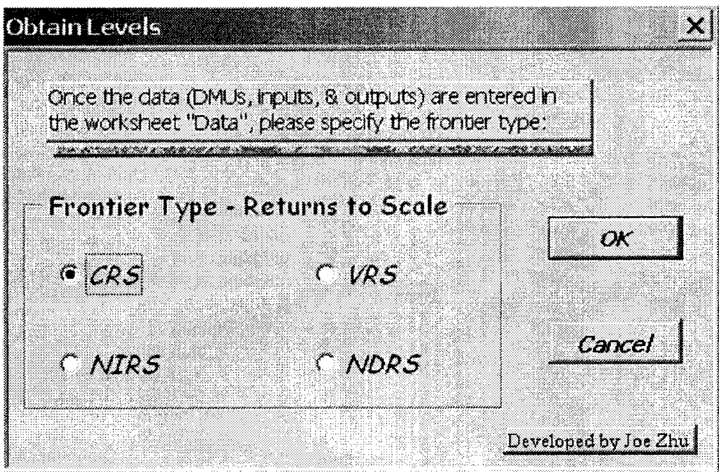

Figure 16-3. Obtain Levels

Once the efficient frontiers are obtained, the context-dependent DEA can be calculated using the "Context-dependent DEA" submenu item as shown in Figure 16-4.

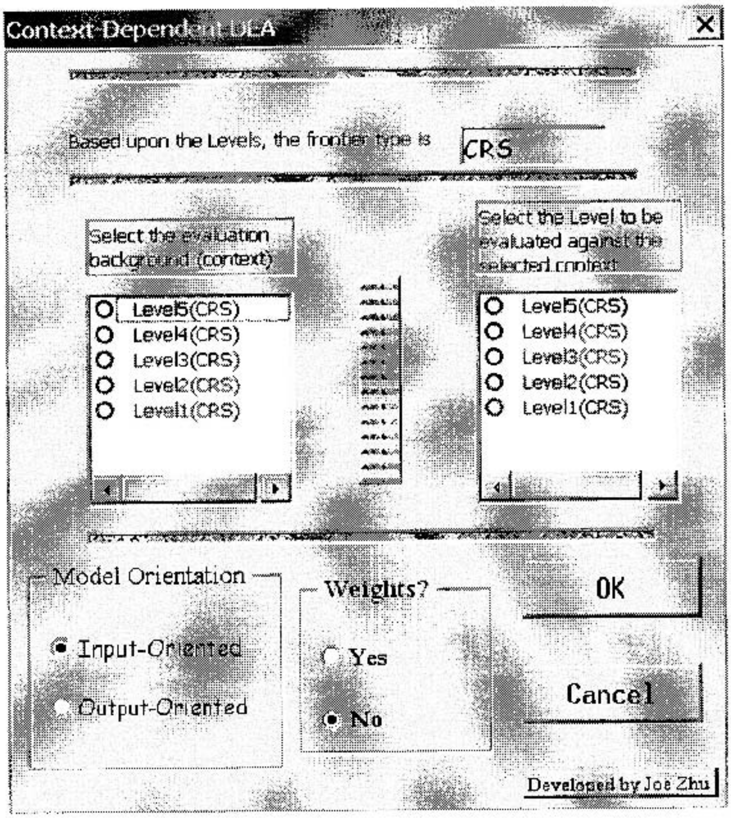

Figure 16-4. [Please provide a caption for this figure]

In Figure 16-4, if one does not wish to specify the weights in model (16.4), for example, then the regular context-dependent DEA model (16.2) is calculated.

The results are reported in the "Context Dependent Result" sheet for the regular context-dependent DEA models, and "Context Dependent VJ" for the models with value judgments. In this sheet, the context-dependent scores are the optimal values to the context-dependent models described in this chapter. To obtain the attractiveness or progress scores, one has to adjust the context-dependent scores based upon Definitions 16.1-16.10.

16.6. APPLICATION

Doyle and Green (1991) benchmarked 37 computer printers using DEA. We revisit their data set by using the newly developed context-dependent DEA. In order to keep the results consistent and comparable with Doyle and Green (1991), we choose price (in US dollars) as the single input. The following features/measures are chosen as outputs: (1) input buffer; (2) mean time between failure (MTBF); (3) 80-column throughput; (4) graphics throughput; (5) sound level and (6) print quality (see Table 16-2).

Table 16-2. Data for the 32 Printers

Printer name	DMU	Price	Input	MTBF	80-column	Graphics	Sound	Print
Epson LQ-500	1	499	8	4000	101	850	72	5
NEC P2200	2	499	8	4000	85	830	72	5
Seikosha SL-	3	549	16	3200	56	451	68	4
Copal WH 6700	4	795	50	4000	102	450	69	3
Epson LQ-850	5	799	38	4000	148	1350	71	7
Printronix P1013	6	895	2	4000	107	683	78	6
Panasonic KX-	7	899	45	4000	107	850	75	7
Brother M-	8	949	32	4000	107	931	72	5
Citizen Tribute	9	949	24	5000	122	917	73	6
ALPS ALQ324	10	995	71	5000	105	562	69	6
Fujitsu DL3400	11	995	24	8000	146	1440	63	7
NEC P7	12	995	50	5000	111	1255	65	6
Sanyo PR-241	13	999	10	8000	90	955	68	6
Dataproducts	14	1099	32	5000	121	687	72	5
Epson LQ-1050	15	1099	48	6000	147	1367	71	7
Facit B3450	16	1245	16	4000	134	1090	72	5
C. Itoh C-715A	17	1295	32	7200	131	1186	74	7
Nissho NP-2405	18	1295	36	6000	139	650	72	7
ALPS P2400C	19	1395	256	6000	146	1000	70	7
Okidata	20	1399	30	4000	184	2400	67	9
Epson LQ-2500	21	1449	40	6000	128	1459	70	6
Fujitsu DL2600	22	1495	80	6000	146	1588	69	8
NEC P5XL	23	1495	40	7000	132	1421	68	7
Radio Shack	24	1599	64	3000	150	465	68	7
AT&T 477	25	1695	80	6000	146	1301	69	7
Hewlett-Packard	26	1695	36	20000	191	542	69	15
Nissho NP-2410	27	1745	54	6000	169	683	71	12
NEC P9XL	28	1795	48	7000	170	1928	68	8
Mannesmann	29	1799	32	4800	205	1069	63	7
C. Itoh C-815	30	1995	42	7200	182	2823	72	10
Fujitsu DL5600	31	2195	24	8000	236	3176	68	12
Japan Dgtl. Labs	32	2495	128	4000	169	497	63	9

There are two kinds of input buffers: standard and optional. Because some printers have zero values for either the standard or optional input buffer, we combine the two scores to give a composite input buffer score so that all scores are positive. The larger the buffer, the more output a computer can transmit to the printer and the sooner the computer is freed for other uses. As stated in Stewart (1988), MTBF (in hours) is a significant specification of a manufacturer's rating of the durability of a printer. The current study does not have access to the MTBF of the following 5 printers: Star Micronics NB24-15, Toshiba P341SL, IBM Proprinter XL24, Star Micronics NB-15 and Toshiba P351SX.

The third and fourth outputs are measures of printing speed in characters per second (cps) which is the document length in bytes divided by the number of seconds to print it. (Higher numbers signify faster performance.) The fifth output is a measure of the noise level (in dBA) where lower numbers are preferable. Based upon Seiford and Zhu (2002), because it is an output measure, we subtract each number from 100 to obtain an adjusted score for the DEA analysis. The last output is a combined quality score for text and graphics quality scores where larger numbers indicate a higher quality. Note that the last four outputs are among the test criteria used by Stewart (1988). Also, based upon Stewart (1988), printers 1 to 13 are in the low price category ($499-$999), printers 14 to 23 are in the middle price category ($1000-$1499), printers 24 to 30 are in the high price category ($1500-$1999) and printers 31 and 32 are in the deluxe price category ($2000-$2499).

By using the DEA model (16.1), we obtain five levels of efficient frontiers. They are,

$$\mathbf{E}^1 = \{DMU_j | \mathrm{j} = 1, 2, 3, 5, 19, 20, 26\}$$

$$\mathbf{E}^2 = \{DMU_j | \mathrm{j} = 4, 7, 10, 11, 12, 15\}$$

$$\mathbf{E}^3 = \{DMU_j | \mathrm{j} = 6, 8, 9, 13, 22, 27, 30, 31\}$$

$$\mathbf{E}^4 = \{DMU_j | \mathrm{j} = 14, 16, 17, 18, 21, 23, 25, 28, 29, 32\}$$

$$\mathbf{E}^5 = \{DMU_j | \mathrm{j} = 24\}$$

It can be seen from the original DEA (CCR) model, seven printers in $\mathbf{E}^1$ are efficient. This result is slightly different from that of Doyle and Green (1991), partly because we treat one of the outputs, sound level, in a different way. Note that three of the six "outstanding buys" selected by Stewart (1998), namely, DMU1 (Epson LQ-500), DMU20 (Okidata Microline 393) and DMU26 (Hewlett-Packard RW480) are in the first-level efficient frontier and the remaining three, namely, DMU4 (Copal WH6700), DMU11 (Fujitsu DL3400) and DMU31 (Fujitsu DL5600) are in the second-level and third-level efficient frontier, respectively. We next discuss the 13 printers in $\mathbf{E}^1$ and $\mathbf{E}^2$ in detail.

Table 16-3. Attractiveness and progress scores for the 13 printers in $\mathbf{E}^1$ and $\mathbf{E}^2$

		Background (Efficient frontier)		
Printer Name	DMU No.	2nd-level first-degree	3rd-level second-degree	4th-level third-degree
Epson LQ-500	1	1.50092③	1.85446③	2.32072②
NEC P2200	2	1.50092③	1.78060④	2.30622③
Seikosha SL-	3	1.51699②	1.85487②	2.28781④
Epson LQ-850	5	1.33046⑤	1.59208⑤	1.83955⑥
ALPS P2400C	19	2.57175①	3.42936①	3.57769①
Okidata	20	1.18545⑥	1.31406⑦	1.59716⑦
Hewlett-	26	1.46755④	1.54022⑥	2.12224⑤
		first-degree**	first-degree	second-degree
Copal WH	4	1.16312③	1.50432①	1.72718①
Panasonic KX-	7	1.13117②	1.28282⑤	1.53604⑤
ALPS	10	1.27868⑦	1.41235②	1.60646③
Fujitsu	11	1.03020①	1.36654③	1.67563②
NEC P7	12	1.19557⑤	1.29240④	1.57736④
Epson LQ-	15	1.22295⑥	1.18763⑥	1.38622⑥

*The number to the right of each score indicates the ranking position.

** This represents progress.

First, by using (16.2) we consider the attractiveness and progress of the fourteen printers when different efficient frontiers are chosen as evaluation contexts. Table 3 gives the results.

The number to the right of each score indicates the ranking position by the attractiveness measure. (① represents the top-rank position.) Note that DMU19 (ALPS P2400C) and DMU4 (Copal WH 6700) are the most attractive printers in the first and second levels, respectively, no matter which evaluation context is chosen. Also, DMU1 (Epson LQ-500) and DMU11 (Fujitsu DL3400) have the second and third ranking positions, respectively.

In fact, for DMUs that are not located on the first or last level of efficient frontier, we can characterize their performance by their attractiveness and progress as shown in Figure 16-5 where the solid circle represents the DMU being evaluated. The most desirable category is the Low Progress – High Attractiveness (LH) and the least desirable category is the High Progress – Low Attractiveness (HL). A high progress indicates that the DMU needs to improve its outputs substantially, and a high attractiveness indicates that the DMU does not have any close competitors. For example, for the printers in $\mathbf{E}^2$, we may categorize (i) Copal WH 6700 (DMU4) and Fujitsu DL3400

(DMU11) as LH, (ii) Panasonic KX-P1524 (DMU7) as LL, (iii) ALPS ALQ324 (DMU10) as HH, and (iv) NEC P7 (DMU12), Epson LQ-1050 (DMU15) as HL.

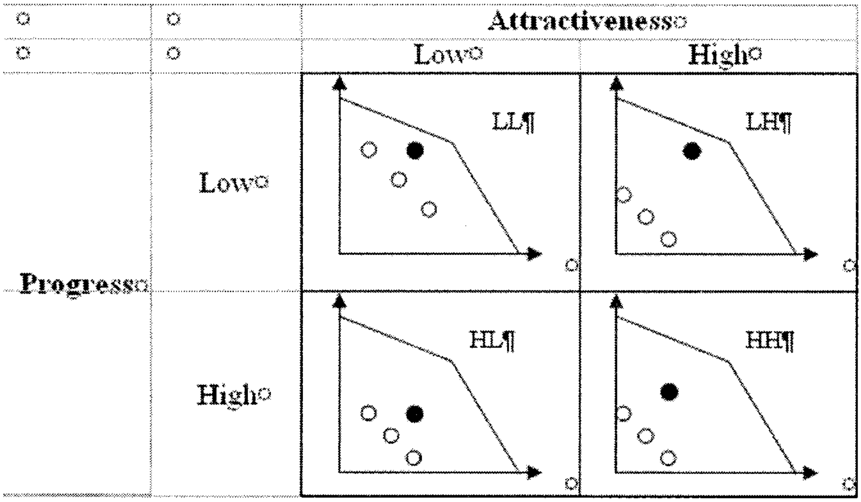

Figure 16-5. Attractiveness – Progress

Next, we consider DMU19 (ALPS P2400C). Note that this printer has the largest input buffer, 256k (the average value of the others is 40k). Thus, the massive input buffer is likely to lead to the large attractiveness score for that printer, and consequently, the attractiveness measure for DMU19 may be biased. Therefore, we need to define some weights, u_r (r = 1,..., 6) to construct the output-oriented VJ attractiveness score by using model (16.4).

Stewart (1998) writes:

> *"Among low-price units, the Epson LQ-500 ($499), the Copal Write Hand 6700 ($795), and the Fujitsu DL3400 ($995) each offer bargain hunters good combinations of speed and quality."*

Thus, if we prefer speed and quality, we specify the following weights where more weight is put on 80-column throughout, graphics throughout and print quality which characterize speed and quality.

Weight-1: u_1 = 0.004, u_2 = 0.003, u_3 = 0.33, u_4 = 0.33, u_5 = 0.003, u_6 = 0.33

Tables 16-4 and 16-5 report the VJ (first-degree) attractiveness scores for the printers in $\mathbf{E}^1$ and $\mathbf{E}^2$, respectively.

Table 16-4. VJ attractiveness scores for the seven printers in $\mathbf{E}^1$ when $\mathbf{E}^2$ is chosen as the evaluation context

Printer Name	DMU No	no weight	weight-1	weight-3
Epson LQ-500	1	1.50092③	1.31529①	1.39505
NEC P2200	2	1.50092③	1.23141②	1.27585
Seikosha SL-80AI	3	1.51699②	1.00125⑦	1.00208
Epson LQ-850	5	1.33046⑤	1.22384③	1.25242
ALPS P2400C	19	2.57175①	1.00255⑥	1.00319
Okidata Microline 393	20	1.18545⑥	1.05441④	1.00078
Hewlett-Packard RW480	26	1.46755④	1.04331⑤	1.06608

*The number to the right of each score indicates the ranking position.

Table 16-5. VJ attractiveness scores for the seven printers in $\mathbf{E}^2$ when $\mathbf{E}^3$ is chosen as the evaluation context

Printer Name	DMU No	no weight	weight-1	weight-2
Copal WH 6700	4	1.50432①	1.00320	1.00431
Panasonic KX-P1524	7	1.28282⑤	1.06673	1.14002
ALPS ALQ324	10	1.41235②	1.00263	1.00347
Fujitsu DL3400	11	1.37561③	1.20117①	1.03405
NEC P7	12	1.29241④	1.01478	1.00407
Epson LQ-1050	15	1.18991⑥	1.08059	1.00399

*The number to the right of each score indicates the ranking position.

It can be seen that DMU1 (Epson LQ-500) and DMU11 (Fujitsu DL3400) are the top-ranked printers in $\mathbf{E}^1$ and $\mathbf{E}^2$, respectively. Note that DMU11 (Fujitsu DL3400) is the top-ranked unit among the inefficient DMUs by the CCR model (see Figure 2). This observation strengthens the conclusion that these two printers are the best ones.

However, DMU4 (Copal WH6700) which is ranked highly by the CCR model does not have a large attractiveness score. When calculating the VJ attractiveness score for DMU4, model (4) identifies DMU8 and DMU9 as the referent DMUs. (The associated optimal lambda values are 0.013 and 0.824, respectively.) Thus, the unattractiveness of DMU4 is due to the presence of DMU8 and DMU9. Note that DMU4, DMU8 and DMU9 are all in the low price category. Hence, DMU8 (Brother M-1724L) and DMU9 (Citizen Tribute 224) could be the potential competitors for DMU4 (Copal WH6700).

It can also be seen that DMU26 (Hewlett-Packard RW480) has a small attractiveness score of 1.04331 although it achieves a top rating in terms of text and graphics quality. Note that our VJ attractiveness measure is based on the situation where inputs are fixed at current levels. Model (16.4) identifies DMU7 (Panasonic KX-P1524) as the referent printer. If we examine the original data for the two printers,

Printer name	DMU No.	Price	Input buffer	MTBF	80-column throughput	Graphics throughput	Sound level	Print quality
Panasonic KX-P1524	7	899	45	4000	107	850	75	7
Hewlett-Packard RW480	26	1695	36	20000	191	542	69	15

we observe that the price of DMU26 almost doubles that of DMU7. Note that DMU7 is in the low price category and DMU26 is in the high price category. However DMU26 does not have a higher value of graphics throughput, and consequently, the presence of DMU7 makes DMU26 less attractive. DMU7 may be a better alternative for DMU26 if one's budget is restricted. In other words, in terms of the price and the printers in $\mathbf{E}^2$, DMU26 (Hewlett-Packard RW480) is not attractive among the seven printers in $\mathbf{E}^1$. This result is consistent with the statement in Stewart (1988, p124): "If you're willing to pay the price, you can definitely find speed and quality in one unit (Hewlett-Packard RW480)". Finally, note that DMU19 dropped to the sixth position in terms of attractiveness ranking.

If quality alone is the consideration, then we choose the following weights:

Weight-2: $u_r = 0.005$ (r = 1, ..., 5) and $u_6 = 0.975$

From the last column of Table 5, we see that the most attractive printer is DMU7 (Panasonic KX-P1524), followed by the DMU11 (Fujitsu DL3400) which were suggested by Stewart (1988) for quality consideration.

If we prefer 80-column throughout and quality, we specify the following weights:

Weight-3: $u_1 = 0.005$, $u_2 = 0.005$, $u_3 = 0.49$, $u_4 = 0.005$, $u_5 = 0.005$, $u_6 = 0.49$

In this case, DMU20 (Okidata Microline 393) is the most unattractive printer among the seven printers in $\mathbf{E}^1$ (see last column in Table 16-4). Stewart (1988) stated "The Okidata Microline 393 ($1399) looks more like a high-price unit in terms of 80-column throughout and quality". In fact, DMUs 11 is in the reference set under model (16.4), i.e., this DMU serves as the evaluation context when measuring the VJ attractiveness of DMU20. In terms of the price, DMU20 obviously does not have the advantage in 80-column throughout and quality.

Table 16-6. Output-specific attractiveness scores for the printers in $\mathbf{E}^1$

Printer name	DMU No.	Input buffer	MTBF	80-column throughput	Graphics throughput	Sound level	Print quality
Epson LQ-500	1	1	1	1.379	1.177	1.509	1.424
Referent Printer		Fujitsu DL3400 (DMU11)					
NEC P2200	2	1	1	1.161	1.149	1.509	1.424
Referent Printer		Fujitsu DL3400 (DMU11)					
Seikosha SL-80AI	3	1	1	1	1	1.165	1.015
Referent Printers		Panasonic KX-P1524 (DMU7) and Fujitsu DL3400 (DMU11)					
Epson LQ-850	5	1	1	1.608	1.320	1.026	1.428
Referent Printer		Fujitsu DL3400 (DMU11) and NEC P7 (DMU12)					
ALPS P2400C	19	2.749	1	1	1	1	1
Referent Printer		ALPS ALQ324 (DMU10) and NEC P7 (DMU12)					
Okidata Microline 393	20	1	1	1	1.185	1	1
Referent Printers		Fujitsu DL3400 (DMU11)					
Hewlett-Packard RW480	26	1	1.468	1	1	1	1.258
Referent Printer		Fujitsu DL3400 (DMU11)					

Finally, we illustrate how to identify which of the six features (outputs) of each printer in $\mathbf{E}^1$ exhibits the leading performance with respect to the printers in $\mathbf{E}^2$. That is, based upon $\mathbf{E}^2$ and the first-degree attractiveness, we determine, for a printer in $\mathbf{E}^1$, (a) the "superior" features that other printers may have difficulties to catch up, and (b) the "noninferior" features for which other printers or their combinations also achieve the same performance level. This analysis provides the manufacturers with information on (i) which features of a printer should be improved to gain a competitive edge, and (ii) the referent printers in $\mathbf{E}^2$ may be viewed as potential competitors.

Let us assume equal weights in model (16.4), i.e., $u_r = 1/6$, $r = 1, \ldots, 6$. Table 16.6 reports the six output-specific attractiveness measures along with the referent printers. (The DEAFrontier software provides the information by Output Changes in the "Context Dependent VJ" sheet.)[1] It can be seen that four printers in $\mathbf{E}^2$ appear in the reference set, of which three are outstanding buys, and in particular, Fujitsu DL3400 (DMU11) almost appears in every reference set. The two outstanding buys in $\mathbf{E}^1$, namely

[1] The scores in Table 16.6 are reciprocals to the "Output Changes" in the "Context Dependent VJ" sheet based upon Definition 16.5.

Okidata Microline 393 (DMU20) and Hewlett-Packard RW480 (DMU26), which are in the high/deluxe price category, do not exhibit good performance in terms of output-specific attractiveness measures. For instance, DMU20, which is the winner (middle price) in graphics tests, only has 1.185 on its graphics throughput, and 1.0 on all other features. DMU26 exhibits good performance only on MTBF and print quality. However, Epson LQ-850 (DMU5) exhibits a good performance based upon many of the output-specific attractive measures. This indicates that if no preference is given to specific output features, this printer may be a good choice in the presence of the outstanding buy, DMU11 (Fujitsu DL3400).

16.7. CONCLUSIONS

Context-dependent DEA is developed to measure the attractiveness and progress of DMUs with respect to a given evaluation context. Different strata of efficient frontiers rather than the traditional first-level efficient frontier are used as evaluation contexts. In the original DEA, adding or deleting inefficient DMUs does not alter the efficiencies of the existing DMUs and the efficient frontier whereas under the context-dependent DEA, such action changes the performance of both efficient and inefficient DMUs. i.e., the context-dependent DEA performance depends on not only the efficient frontier, but also the inefficient DMUs. This change makes DEA more versatile and allows DEA to locally and globally identify better options. Value judgment is incorporated into the context-dependent DEA through a specific set of weights reflecting the preferences over various output (or input) measures. In particular, the attractiveness measure can be used to (i) identify DMUs that have outstanding performance and (ii) differentiate the performance of DEA efficient DMUs.

The application of comparing computer printers illustrates that in-depth information can be obtained by the context-dependent DEA when compared to the results obtained from the original DEA method. Context-dependent DEA identifies the most attractive printer among the outstanding buys located at two different levels of efficient frontiers. It also identifies the most attractive printer in terms of individual features, e.g., speed and quality. The method uncovers better options and prescribes possible improvement when a specific printer is rated as inefficient by the original DEA model. With a restricted budget, the DEA-efficient printers may not necessarily be the best choice. In our application, we are able to identify better alternatives. In addition, with a sensitivity analysis of weights, one could determine allowable weight ranges to be specified by users or experts.

REFERENCES

1. Charnes A., W.W. Cooper, E. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2, 429-441.
2. Doyle J.R. and R.H. Green (1991), Comparing products using data envelopment analysis, *OMEGA* 19, 631-638.
3. Simonson I. and A. Tversky (1992), Choice in context: Tradeoff contrast and extremeness aversion, *Journal of Marketing Research* 29, 281-295.
4. Seiford L.M. and J. Zhu (2002), Modeling undesirable factors in efficiency evaluation, *European Journal of Operational Research* 142, 16-20.
5. Stewart G.A. (1998), The best of the 24-pin printers, *BYTE* April, 118-128.
6. Thompson R.G., L. Langemeier, C.T. Lee, E. Lee, and R.M. Thrall (1990), The role of multiplier bounds in efficiency analysis with an application to Kansas farming, *J. of Econometrics* 46, 93-108.
7. Tversky A. and I. Simonson (1992), Context-dependent preferences. *Management Sciences* 39, 1179-1189.
8. Zhu J. (1996), Data envelopment analysis with preference structure, *J. Opl Res Soc* 47, 136-150.
9. Zhu J. (2002), Quantitative models for performance evaluation and benchmarking: Data envelopment analysis with spreadsheets and DEA excel solver, Boston: Kluwer Academic Publishers.

This chapter is based upon Seiford, L.M. and Zhu, Joe, 2003, Context-dependent data envelopment analysis: measuring attractiveness and progress. OMEGA, Vol. 31, Issue 5, 397-480, with permission from Elsevier Science.

Chapter 17

EVALUATING POWER PLANT EFFICIENCY

Hierarchical Models

17.1. INTRODUCTION

In many problem settings that potentially lend themselves to analyses via DEA, there are identifiable groups or clusters of DMUs, whose impacts should be captured in the analysis. One form of grouping has been examined by Banker et. al. (1986), where the idea of categorical variables was discussed. Such variables allow for a comparison of any DMU with those in its own category and in those categories *below* it. Categorical variables generally apply in those situations where there is a natural nesting of the groups of DMUs. For example, in evaluating a set of banks, if the banks are arranged in increasing order according to the sizes of the towns or cities in which they are located, then categorical variables can be used to represent this size component, and banks in a given population category will be compared only to DMUs in this same category and to those in smaller population categories.

In many situations, however, where there is a grouping phenomenon present, categorical variables do not provide an appropriate structure for analysis. Consider the problem of evaluating DMUs such as hospitals in different parts of the country. Here, grouping may take several forms. First, in countries such as Canada or the United States, there may be jurisdictional considerations, e.g., state or provincial regulations can have budgetary or legislation implications for the hospitals. In Canada, for example, health care is under provincial rather than federal jurisdiction. Second, there may be different categories of medical units - extended care facilities, convalescent units, surgical units, and so on. Clearly, these DMUs do not form anything resembling a homogeneous set, making it necessary to address the group elements of the problem. At least two issues must be examined in the context of such problems:

Issue 1: There are both DMU (e.g., hospital) level factors and group (e.g., all extended care facilities versus all surgical facilities) level factors which should be dealt with in their proper settings;

Issue 2: We want not only to identify a measure of efficiency for each individual DMU (hospital), but also for each identified group of units. How do hospitals (as a group) in one jurisdiction compare, in an efficiency sense, to those in another jurisdiction? Do extended care facilities perform differently than surgical facilities?

In the following sections we examine the problem of efficiency evaluation when grouping of DMUs is a consideration. The discussion is based on the articles by Cook et al. (1998) and Cook and Green (2004). In Section 17.2 we present a problem setting where both individual DMU and group evaluation arise. The case illustrates two types of grouping - *hierarchical* grouping and grouping on *levels*. Section 17.3 presents appropriate model structures for evaluating group and individual DMU efficiency in a hierarchy. In particular we discuss a procedure for adjusting ratings of DMUs at any given level in a hierarchy to reflect ratings of groups of those units at levels higher up in the structure. In Section 17.4 we examine efficiency within groups on a level and develop a procedure for combining different efficiency ratings for a given DMU. In Section 17.5 the models are illustrated through an analysis of the application discussed in section 17.2. In Section 17.6 the power plant evaluation problem is re-examined using the multicomponent concepts presented in Chapter 6. This arises from the need to deal with output shared among power units within a grouping. Section 17.7 illustrates the concepts using data similar to that found in Section 17.5. Conclusions and further directions follow in Section 17.8.

17.2. HIERARCHICAL STRUCTURES: POWER PLANTS

Ontario Hydro (now called Ontario Power Generation) is a crown corporation supplying electric power to both domestic and foreign markets in the northern USA. Two classes of power units or plants are managed under Hydro's jurisdiction – nuclear and thermal. While the number of nuclear units is relatively small, a total of 40 such units of varying ages, capacities, fuel types and so on are operated by the corporation. These latter will be the setting for the analysis of section 17.5.

The standard measure of productivity used by management is the ratio of total annual expenditure (operating, maintenance and administration) to total energy produced, in megawatt hours per year. While it is the case that the total power production is a principal *output* of the operation, and is certainly the most convenient and readily available indicator of productive capability, management is interested in other, related indicators as well. What may be missing in this simplistic measure of productivity is a consideration of those factors that reflect management's skill. To a great extent, a power unit's efficiency measure should reflect the quality of maintenance that keeps it operating, and the abilities of management in charge of that maintenance. At least two types of other outputs should be considered, namely *outages* and *deratings*.

An outage is a situation in which a unit is shut down, hence it is not generating electric power. Types of outages include:

- planned outage, which is scheduled downtime (usually for major overhauls);
- maintenance outage, a form of scheduled down time, but for more minor, i.e., routine maintenance;
- forced outage, which is unscheduled and generally caused by equipment failure, environmental requirements, or other unforeseen incidents. There is generally some prior warning for this type of shutdown, and some delay is possible;
- sudden outage, which is a forced outage with no prior warning.

While it can be argued that operating hours essentially capture all forms of outages, it must be recognized that there is a difference between taking a unit out of service on a scheduled basis at non peak times, versus sudden brownouts or blackouts. The latter ignite public opinion, interrupt business operations, and generally reflect negatively on the organization. Thus, such outages should play a direct role in any measure of efficiency.

A derating is a *reduction* in unit capacity, where the operation may, for a number of reasons, operate at only a fraction (e.g., 75% or 50%) of its available (full) capacity. Breakdowns in coal belts, pulverizers or rollers (of which there are several operating in any plant) is a primary cause of such forced deratings. Environmental restructions, in particular SO_2 emissions, can limit the extent to which a plant can operate at full capacity. Furthermore, such restrictions will often apply to a group of units (e.g., at a given geographical location).

As with outages, there are several forms of deratings, some of which are beyond the control of management and which have nothing to do with maintenance quality (e.g., grid or transmission network load restrictions), while others are a clear reflection of maintenance quality, such as equipment failures.

As with outputs, inputs should include several factors. In addition to expenditures, factors such as plant *age* and *available but not operating time* (ABNOT) should play a role as well. The latter factor (ABNOT) is the time during which the plant is able to operate, but for reasons beyond managements control (such as SO_2 restrictions), the plant is not running.

Grouping is a natural phenomenon here. Plants can be grouped by size or capacity, by geographical location, and so on. It is this necessity to view problems from a grouping and hierarchical perspective that we examine herein.

17.3. MODELS FOR EVALUATING PLANT HIERARCHIES

The power plant application discussed in the previous section provides an example of what might be termed a *pure hierarchy*. The basic DMU is the power unit. These 40 units are naturally clustered into 8 plants.

17.3.1 The Two-Level Hierarchy

For simplicity of presentation in this subsection we assume there are only two levels in the hierarchy. Let the level 1 (power units) vectors of inputs and outputs be denoted $X(1), Y(1)$ respectively, with $v(1), \mu(1)$ representing the appropriate multipliers in the input orientation version of the CCR (Charnes et al. 1978) model.

In the normal case where we are interested only in a level 1 (power unit) analysis of efficiency, the "multiplier" form of the CCR model is:

$$\max \mu^T(1)Y_o(1) \tag{17.1a}$$

$$\text{subject to:} \quad v^T(1)X_o(1) = 1 \tag{17.1b}$$

$$\mu^T(1)Y_j(1) - v^T(1)X_j(1) \leq 0, \qquad j \varepsilon J \tag{17.1c}$$

$$\mu(1), v(1) \geq \varepsilon, \tag{17.1d}$$

where J is the set of DMUs under consideration. Suppose, however, that we want, in addition, to evaluate the relative efficiencies of the 8 plants into which the 40 units are grouped. Clearly, one approach might be simply to evaluate each DMU relative to the entire set of 40 units as indicated above, (hence J would represent the entire set of power units), and use the average of the ratings for those units within any plant as representative of the standing of that plant. While it is difficult to argue that such an approach is wrong, it does possess some undesirable aspects. First, those factors that apply at the group level (level 2) are not represented (or at least not represented appropriately) in level 1. Second, and as indicated above, it would seem more appropriate at level 1 to evaluate a DMU relative to those DMUs in the same group only. In this case, J in (17.1c) above would refer to those units in a specific plant, whereupon, those factors which distinguish the groups (plants) can be omitted from the level 1 evaluation, and can more properly be applied at level 2. If this is done, then averaging within a plant does not help at all to understand the relative standings of the level 2 DMUs.

An alternative approach for evaluating efficiency at both levels 1 and 2, is to treat the level 2 groups themselves as decision making units, using a combination of the group-specific factors, and factors which emerge from level 1. The use of level 1 factors at level 2 may involve some form of aggregation as will be discussed in the next section.

For notational purposes define

K - the number of *groups* of level 1 DMUs, hence K is the number of DMUs at level 2;

k - a subscript representing a DMU at level 2;

j_k - a subscript representing a level 1 DMU that belongs to group k;

$Y_{kj_k}(1), X_{kj_k}(1)$ - level 1 outputs and inputs;

$Y_k^1(2), X_k^1(2)$ - those level 2 outputs and inputs that are aggregates of factors that are used to evaluate level 1 DMUs;

$Y_k^2(2), X_k^2(2)$ - those outputs and inputs used at level 2 that distinguish the K groups, and which were not used at level 1.

Let $\nu(1), \mu(1)$ and $\nu(2), \mu(2)$ denote the level 2 multipliers to be associated with $X_k^1(2), Y_k^1(2)$ and $X_k^2(2), Y_k^2(2)$ respectively. It is noted that $\nu(1), \mu(1)$ are the same multipliers as used in level 1, as will be explained below.

In performing the analysis within a general model framework we make the following assumptions:

(a) When the "DMU" under consideration is a level 1 unit, we want to ensure that it is evaluated only relative to those units in the same group, hence DMUs in other groups should be excluded or *disengaged* from the constraint set;

(b) Level 2 DMUs (groups) should not interfere with, hence should be disengaged from, level 1 analyses;

(c) Level 1 DMUs should be included or *engaged* in the analysis of level

2 units.

Assumption (c) above is invoked with the argument that multipliers $\nu(1), \mu(1)$ when applied at level 2 should also be feasible when applied to any level 1 units within the groups under consideration. Specifically, since the efficiency of a given group i at level 2 should be related to the efficiencies of that group's members, then the multipliers $\nu(1), \mu(1)$ should be such that when applied to each member of the group, the ratio for that member should not exceed unity.

To accommodate the above considerations we propose the following general model. When applied to a level 2 DMU, the model would take the form:

$$\max e_o = \mu^T(1)Y_0^1(2) + \mu^T(2)Y_0^2(2) \quad (17.2a)$$

subject to:

$$v^T(1)X_0^1(2) + v^T(2)X_0^2(2) + Mw(2) = 1 \quad (17.2b)$$

$$\mu^T(1)Y_k^1(2) + \mu^T(2)Y_k^2(2) - v^T(1)X_k^1(2) - v^T(2)X_k^2(2) - w(2) \le 0 \;\; k = 1,...K \quad (17.2c)$$

$$\mu^T(1)Y_{kj_k}(1) - v^T(1)X_{kj_k}(1) - w_k(1) \le 0, j_k \varepsilon J_k, k = 1,...,K \quad (7.d)$$

$$w(2) - w_k(1) \ge 0, \qquad k = 1,...,K \quad (17.2e)$$

$$\mu(1), \mu(2), v(1), v(2) \ge \varepsilon \quad (17.2f)$$

$$w_k(1), w(2) \ge 0, \qquad \forall k \quad (17.2g)$$

When applied to a level 1 DMU, the model would take the form

$$\max e_0 = v^T(1)Y_0(1) \quad (17.2a')$$

subject to:

$$v^T(1)X_0(1) + Mw_0(1) = 1 \quad (17.2b')$$

(17.2c)-(17.2g)

Here M denotes a large positive number. In (17.2d), J_k denotes the index set of level 1 DMUs in group k (plant k). The notation $Y_0^1(2)$ in (17.2a) denotes the type 1 output (an aggregate of level 1 outputs) used at the second level and for a particular DMU $k = "0."$. In (17.2a') the notation $Y_0(1)$ denotes the output at level 1 for a particular DMU and group $(j_k, k) = "0."$ The variables $w_k(1), w(2)$ are introduced to include or exclude certain DMUs from the analysis. In reference to the above discussion, these are referred to as *engagement variables*. It is noted that in (17.2b'), $w_0(1)$ refers to the particular level 2 groups $k = "0"$ in which the

DMU under evaluation lies at the time. So, for example, all DMUs in a particular group k will be assigned the same variable $w_0(1)$.

In reference to assumptions (a), (b) above, we prove the following theorem:

Theorem 17.1: In the evaluation of any level 1 DMU $j_{k_0} \varepsilon k_o$, only DMUs within the same group (k_0) as that DMU will be engaged. All other level 1 groups and all level 2 DMUs are disengaged.

Proof: Only the particular group $k = "0"$ in which the level 1 DMU under consideration lies, has its engagement variable $(w_0(1))$ involved in constraint (17.2b'). This variable will be forced to zero, otherwise the objective function value of (17.2a') will equal zero. Furthermore, since all other engagement variables are free to assume the most favorable values possible (from the perspective of the DMU under evaluation), then all $w_k(1)$ (except for $w_0(1)$) and $w(2)$ will assume values large enough to render redundant all constraints in (17.2c), as well as all constraints in (17.2d) corresponding to those $J_k, k \neq "0."$ Since constraints (17.2e) are also redundant, the result follows.

QED

From this theorem it follows as well that when a level 2 DMU is under evaluation, the engagement variable $w(2)$ will be forced to zero (hence *engaging* all level 2 DMUs). By virtue of constraints (17.2e), all $w_k(1) = 0$ as well, hence engaging all level 1 DMUs, thereby verifying assumption (c).

17.3.2 Efficiency Adjustments in a Hierarchy

In Section 17.5 we present an analysis of the efficiencies of power plants and groups of plants. One issue that arises in such multi level analyses has to do with adjustments in DMU efficiencies at one level to account for scores assigned at a higher level. Specifically, the scores achieved by individual DMUs (e.g., level 1) are measured only against others in the same group. To adjust these to reflect the standings of the groups themselves, it is necessary to merge the scores at these two levels in some reasonable manner. We describe a three step procedure to bring about the desired adjusted ratings.

Step 1: (Remove inter-group noise)

Scale the level 1 ratings by dividing each rating e_{kj_k} in group k by the average of the group k ratings. Specifically, define

$$f_{kj_k} = e_{kj_k}/\overline{e}_k \text{ where } \overline{e}_k = (\sum_{j_k \varepsilon J_k} e_{kj_k})/|J_k|$$

where $|J_k|$ denotes the cardinality of J_k. Since the level 2 ratings are intended to account for any inter-group differences, this transformation is intended to remove any differences (noise) among the groups that are not level 2 - related. See Property 17.2 below and explanation following it.

Step 2: (Introduce level 2 adjustment)

Adjust the scaled ratings f_{kj_k} by multiplying them by the level 2 (group) ratings e_k. That is, define

$$g_{kj_k} = f_{kj_k} \times e_k.$$

Step 3: (Adjust to [0,1] scale)

Further adjust the step 2 ratings g_{kj_k} to ensure that the maximum level 1 rating is unity. Specifically, we want to adjust the g_{kj_k} ratings to the form

$$h_{kj_k} = g_{kj_k} \times R$$

where R is such that $h_{kj_k} \leq 1$, and $\max_{k,j_k}\{h_{kj_k}\} = 1$. Hence $R = \min_{k,j_k}\{1/g_{kj_k}\}$.

The final adjusted ratings therefore have two important properties:

Property 17.1: All level 1 ratings $h_{kj_k} \leq 1$, with at least one $h_{k_o j_{k_o}} = 1$.
Property 17.2: The averages of the ratings $\overline{h}_k$ within the K groups are such that $\frac{\overline{h}_{k_1}}{\overline{h}_{k_2}} = \frac{e_{k_1}}{e_{k_2}}$.

The latter property captures the fact that the final adjusted ratings not only represent the standing of DMUs (e.g., power units) within their own group k (plant), but also reflect their standing relative to DMUs in other groups. That is, if the rating e_{k_1} of one group k_1 is, for example, only 80% of the rating e_{k_2} of another group k_2, then the averages for the DMUs in the two groups, namely $\overline{h}_{k_1}$ and $\overline{h}_{k_2}$, have this same property.

17.3.3 The Multi level Hierarchy

The model (17.2a)-(17.2g) can be generalized to the case of an L-level hierarchy. We assume that the outputs and inputs used at any level ℓ are aggregates of $\ell - 1$ level factors together with any additional factors that distinguish the groups at the ℓ th level. We introduce the following notation:

ℓ - subscript representing a level in an L-level hierarchy;

K_ℓ - the number of DMUs at the ℓ th level;

k_ℓ - a subscript representing a DMU at level ℓ;

j_{k_ℓ} - a subscript representing a DMU at level $\ell-1$ that lies within a group k_ℓ (that is, within a DMU k_ℓ at the next level up in the hierarchy);

J_{k_ℓ} - the subset of DMUs j_{k_ℓ} at level $\ell-1$ that lie within group k_ℓ;

$\{Y^m_{k_\ell j_{k_\ell}}(\ell-1), X^m_{k_\ell j_{k_\ell}}(\ell-1)\}_{m=1}^{\ell-2}$ - those outputs and inputs used at level $\ell-1$ that are aggregates of factors used for analysis of DMUs at lower levels $m \le \ell-2$. The subscript k_ℓ refers to the particular $\ell-1$ level group (i.e. ℓ th level DMU), and j_{k_ℓ} to a DMU within that group;

$Y^{\ell-1}_{k_\ell j_{k_\ell}}(\ell-1), X^{\ell-1}_{k_\ell j_{k_\ell}}(\ell-1)$ - those outputs and inputs at level $\ell-1$ that distinguish the DMUs at that level, and which were not used at any lower level;

$w_{k_\ell}(\ell-1)$ - denotes the engagement variables applicable at level $\ell-1$. These distinguish the groups at this level.

$w(L)$ - denotes the engagement variable applicable at level L.

The model, when applied at the $\ell-1$ level then takes the form:

$$\max e_o = \sum_{m=1}^{\ell-1} \mu^T(m) Y_0^m(\ell-1) \tag{17.3a}$$

subject to:

$$\sum_{m=1}^{\ell-1} \nu^T(m) X_0^m(\ell-1) + M w_0(\ell-1) = 1 \tag{17.3b}$$

$$\sum_{m=1}^{L} \mu^T(m) Y_j^m(L) - \sum_{m=1}^{L} \nu^T(m) X_j^m(L) - w(L) \le 0, j = 1,\dots,K_L \tag{17.3c}$$

$$\sum_{m=1}^{\ell-1} \mu^T(m) Y_{k_\ell j_{k_\ell}}(\ell-1) - \sum_{m=1}^{\ell-1} \nu^T(m) X_{k_\ell j_{k_\ell}}(\ell-1) - w_{k_\ell}(\ell-1) \le 0,$$

$$\ell = 2,\dots,L, k_\ell = 1,\dots,K_\ell, j_{k_\ell} \varepsilon J_{k_\ell} \tag{17.3d}$$

$$w(L) - w_{k_L}(L-1) \ge 0, k_L = 1,\dots,K_L \tag{17.3e}$$

$$w_{k_\ell}(\ell-1) - w_{k_{\ell-1}}(\ell-2) \ge 0, k_\ell = 1,\dots,K_\ell, k_{\ell-1} \varepsilon I_{k_\ell} \tag{17.3f}$$

$$\mu(m), \nu(m) \ge \varepsilon, m = 1,\dots,\ell-1 \tag{17.3g}$$

where I_{k_ℓ} is comprised of those sets of DMUs at level $\ell-2$ that make up the k_ℓ th set at level $\ell-1$.

As with the 2-level problem discussed earlier, the engagement variables $w_{k_\ell}(\ell-1)$ act to include or exclude sets of DMUs as the analysis proceeds. With regard to adjustments to ratings, a similar procedure could be applied here by starting at the top level L in the hierarchy to bring about alterations to the ratings at level L-1. Then, apply these adjusted ratings to alter the L-2 level ratings, and so on.

In this subsection we have examined the problem of evaluating DMUs and groups of DMUs which appear in the form of a hierarchy. In the following subsection this idea is extended to look at the alternative groupings of DMUs on the same level.

17.4. GROUPING ON LEVELS

The power plant application discussed above is a prime example of a pure hierarchy in that DMUs are grouped at each level according to a single attribute - in this case a jurisdictional or geographical attribute. In Section 17.5 we analyze the efficiency of the set of power plants and groupings thereof. In this case, the problem of efficiency evaluation seems to invite a 2-level analysis, in that plants can be grouped by a number of different attributes - capacity, geographical location, fuel type and so on. All these factors can be judged as level 2 attributes, although admittedly one can conceive of very complex mixes of these. One could, for example, group plants at the second level according to geographical location, then at a 3rd level group locations by capacity, assuming, of course, that only one capacity of plants exists at a given location. In the present example, this is not exactly the case. Of course, if at the third level we attempted to group by capacity, regardless of the location, then the hierarchical structure is destroyed. Groups at one level would be broken apart when going to the next level.

In the following subsection we will consider grouping only at one level (level 2 in the case of the power plants), and according to multiple attributes. If we wish to have plants at level 1 evaluated strictly within the groups that will form the DMUs at level 2, it would appear that multi attribute grouping implies simply replicating model (17.2a)-(17.2g) as many times as there are attributes. Suppose, for example, that we wish to group plants in two ways: (1) geographical and (2) according to capacity. The most practical approach would appear to be to run this model once for each type of grouping. This would lead to two sets of efficiency ratings. While an elaborate model with engagement variables can easily be formulated, there would seem to be no practical advantage in doing so.

17.4.1 Deriving an Aggregate Rating

The issue of grouping on a level according to a number of different attributes gives rise to the problem of how to derive some form of overall rating for a DMU. Suppose, for example, that plants are grouped by geographical location. A given plant j, when evaluated in a DEA manner, will be compared to other plants within the same group (at the same location). The number of other plants in that group and the efficiencies of those other plants will, of course, influence the score that j receives. When evaluated according to some other grouping attribute such as capacity, plant j will, in all likelihood, receive a different score. The problem then is how to view the aggregate or overall standing of j, given the different ratings for j that arise out of this multi attribute analysis.

One approach to this problem of deriving an overall efficiency measure is to introduce an importance multiplier on the i th attribute. To formalize this, assume there are I attributes, hence I different grouping types, and let e_{ij} denote the efficiency rating received by DMU j when viewed in terms of the grouping created by the i th attribute. Let α_i denote the weight or importance to be accorded attribute $i \varepsilon I$. The α_i may either be supplied weights or may need to be determined (discussed below). Using these multipliers, we define the aggregate efficiency of DMU j to be:

$$e_j = \sum_{i \varepsilon I} \alpha_i e_{ij}$$

In the event that the α_i are decision variables, there may or may not be information available as to appropriate values for these variables. In any event, and in the spirit of general DEA, one approach to deriving an aggregate rating for DMU j_o is to determine $\{\alpha_i\}$ through the optimization procedure

$$e^*_{j_o} = \max\ e_{j_o} = \max \sum_{i=1}^{I} \alpha_i e_{ij_o} \tag{17.4a}$$

subject to:

$$\sum_{i=1}^{I} \alpha_i e_{ij} \leq 1, j \varepsilon J \tag{17.4b}$$

$$\alpha = (\alpha_i, \ldots \alpha_I) \varepsilon \Phi, \tag{17.4c}$$

where Φ defines the available information on the $\{\alpha_i\}$. Constraints (17.4b) bound the problem by requiring that the aggregate efficiency for each DMU not exceed 1.

One minimal set of restrictions on the α_i might be an ordinal ranking of the attributes. Suppose, for example, that the set of attributes consist of:

(b) geographical location,
(c) capacity,
(d) age,
(e) fuel type used.

Furthermore, assume that these attributes can be prioritized in order of importance to the organization (the utility company). With no loss of generality, assume that the most important attribute is geographical location, followed by capacity, then age, and finally fuel type. In notational terms, this would imply that $\alpha_1 > \alpha_2 > \alpha_3 > \alpha_4$. Introducing an infinitesimal ε, Φ may then be defined in this case by

$$\Phi = \{\alpha = (\alpha_1, \ldots, \alpha_I) \mid \alpha_i - \alpha_{i+1} \geq \varepsilon, i = 1, \ldots, I-1; \alpha_I \geq \varepsilon\} \qquad (17.4d)$$

The idea of ordinal relations among multipliers in DEA was discussed in Ali, Cook & Seiford [1991] and Golany [1988]. A somewhat similar structure appears in Cook, Kress and Seiford [1996] in the context of incorporating ordinal data within the DEA framework. Clearly, problem (17.4a)-(17.4d) is a set of J linear problems with each yielding a best or most efficient aggregate evaluation for the DMU j_o under consideration. One possible drawback to this approach is the fact that a different set of $\{\alpha_i\}$ will arise from each of the J optimizations. This, of course, can be a *general* criticism of the DEA approach.

17.4.2 A Common Set of Multipliers

If it is desirable to obtain a single or common set of multipliers $\{\alpha_i\}$, one approach to use in this particular instance is to determine the largest value of ε for which a feasible set of α_i exists. Specifically, solve the *single* optimization problem:

$$\varepsilon^* = \max \varepsilon \qquad (4e)$$
$$\text{subject to (17.4b)-(17.4d)}$$

The set of α_i that are optimal in this problem provide a means of evaluating all DMUs on a common basis. The essence of this approach is that the minimum extent to which we distinguish or discriminate between the importance measures (α_i) attached to the various criteria is maximized.

17.4.3 Multiple Rankings of Attributes

In the above it is assumed that an overall *single* rank ordering of the attributes in I is at hand. This ordering is intended to express the relative importance of the various grouping mechanisms (geographical location, capacity, etc.). In some situations it may be necessary to ask the question "importance in what sense?" If environmental considerations are paramount, the above rank ordering which places geographical location first in importance may be appropriate. On the other hand, if new technology for powering the plants (new fuel types, e.g.,) is an issue, then the attribute 'fuel type' used may rank in first place. Therefore, multiple rankings of the attributes may be in order.

To formalize this concept, assume that Q *ranking vehicles* or mechanisms are to be considered. Let α_i^q denote the importance or weight to be given to attribute $i\varepsilon I$ when viewed from the perspective of ranking vehicle $q\varepsilon Q$. Furthermore, let the *decision variable* β_q represent the weight to be given to vehicle q. While various types of restrictions could be imposed on the β_q, we assume here that only positivity constraints are imposed, i.e. $\beta_q \geq \varepsilon$ for all q. If a rank ordering on the α_i^q is now imposed relative to each q, then Q feasible regions $\{\Phi_q\}_{q=1}^{Q}$ would be defined. Specifically, define

$$\Phi_q = \{\alpha^q = (\alpha_1^q, \alpha_2^q, \ldots, \alpha_I^q) \alpha_{i_\ell}^q - \alpha_{i_{\ell+1}}^q \geq \varepsilon, \ell = 1, \ldots, I-1; \alpha_{i_I}^q \geq \varepsilon\}, \quad (17.4f)$$

where $\alpha_{i_\ell}^q$ denotes the ℓ th ranked attribute from the point of view of the q th ranking vehicle. Following the logic of problem (17.4a)-(17.4c), an aggregate efficiency rating for DMU j_o could then be determined by solving the J problems:

$$e_{j_o}^* = \max e_{j_o} = \max \sum_{q=1}^{Q} \sum_{i=1}^{I} \beta_q \alpha_i^q e_{ij_o} \quad (17.5a)$$

subject to:

$$\sum_{q=1}^{Q} \sum_{i=1}^{I} \beta_q \alpha_i^q e_{ij} \leq 1, j\varepsilon J \quad (17.5b)$$

$$\alpha^q = (\alpha_1^q, \ldots, \alpha_I^q) \varepsilon \Phi_q, q = 1, \ldots, Q. \quad (17.5c)$$

$$\beta_q \geq \varepsilon, q = 1, \ldots, Q \quad (17.5d)$$

Problem (17.5a)-(17.5d), unlike the earlier single ranking vehicle formulation, is nonlinear with the product of the β_q and α_i^q. This formulation can be transformed to an equivalent linear structure, however, through a simple change of variables. That is, define

$$\delta_{qi} = \beta_q \alpha_i^q,$$

and note that the constraints $\alpha^q_{i_\ell} - \alpha^q_{i_{\ell+1}} \geq \varepsilon$ and $\alpha^q_{i_I} \geq \varepsilon$ can be replaced (through multiplication by β_q on both sides of the inequality) by $\delta_{qi_\ell} - \delta_{qi_{\ell+1}} \geq \varepsilon\beta_q$, and $\delta_{qi_I} \geq \varepsilon\beta_q$. Problem (17.5a)-(17.5d) is then equivalent to the linear problem:

$$e^*_{j_o} = \max e_{j_o} = \max\sum_{q=1}^{Q}\sum_{i=1}^{I}\delta_{qi}e_{ij_o} \quad (17.6a)$$

subject to:

$$\sum_{q=1}^{Q}\sum_{i=1}^{I}\delta_{qi}e_{ij} \leq 1, j\varepsilon J \quad (17.6b)$$

$$\delta_{qi_\ell} - \delta_{qi_{\ell+1}} - \varepsilon\,\beta_q \geq 0, \ell = 1,\ldots,I-1; q = 1,\ldots,Q \quad (17.6c)$$

$$\delta_{qi_I} - \varepsilon\,\beta_q \geq 0, q = 1,\ldots,Q. \quad (17.6d)$$

That is, given an optimal solution $(\delta^*_{iq}, \beta^*_q)$ to (17.6a)-(17.6d), then $\alpha^{q*}_i = \delta^*_{iq}/\beta^*_q$ and β^*_q constitute an optimal solution to (17.4f)-(17.4j), due to the fact that all δ^*_{iq}, β^*_q are strictly positive.

In certain situations the ranking vehicles referred to above may take the form of opinions offered by a set of Q voters (e.g. managers). That is, the relative importance of the I grouping attributes may be a matter of opinion, hence model (17.5a)-(17.5d) (and therefore (17.6a)-(17.6d)) is intended to derive a rating which takes into consideration the various opinions (rankings) offered.

Clearly the earlier comments regarding a common set of weights applies in the present situation as well.

In the following section an application is presented which illustrates some of the model structures presented in this and the previous section.

17.5. EFFICIENCY ANALYSIS OF POWER PLANTS: AN EXAMPLE

Earlier a description was given of a problem setting involving thermo generating plants, wherein it was argued that efficiency should be viewed in terms of a set of outputs and inputs. Table 17-1 shows the number of thermal units operating at each of 8 locations, two of which (Plant 4(1) and Plant 7(1)) are each broken down into two groups for a total of 10 groupings. Given also are the construction dates, fuel types and capacities in megawatt hours.

In the analyses of the plants, two levels were examined, namely, the individual power unit level (level 1) and a second level where plants are grouped in various ways. Two forms of analyses were carried out:

(1) a hierarchical analysis at the two levels, where plants are grouped in level 2 by location;
(2) analysis of efficiency on a level where, with different types of grouping, it is necessary to deal with several ratings for a given DMU.

Table 17-1. Thermal Plants

Location	# units	Age Range	Fuel Utilized	Size (MWH)
Plant 1	8	1971-72	U.S. Bit. Coal & Western Cdn. Coal	500
Plant 2	8	1968	U.S. Bit. Coal	300
Plant 3	4	1970	U.S. Bit Coal	500
Plant 4 (1)	1	1964-66	U.S. Bit Coal	100
Plant 4 (2)	2	1974-75	Liquid Bit	150
Plant 5	4	1974	Oil	500
Plant 6	1	1978	Lignite Bit. Coal	200
Plant 7 (1)	4	1956	Gas/Coal	100
Plant 7 (2)	4	1960	Gas/Coal	200
Plant 8	4	1952	U.S. Bit. Coal	50

Table 17-2 displays the raw data for the 40 plants under analysis.[1] Shown are three outputs and two inputs. These outputs and inputs are defined as follows:

Outputs

- OPER - a function of equivalent full capacity operating hours. This factor accounts for the fact that when operating at less than 100% capacity (e.g. if the unit is derated to 50% capacity), the operating hours during this period are prorated. To bring the scale of values for the units of measurement within the range of the scales used for other factors, we apply a scaling factor of $\frac{1}{10}$, i.e. OPER = $\frac{1}{10}\times$ full capacity operating hours.
- OUT — a function of the number of forced and sudden outages.

OUT=N-K(# forced outages + # sudden outages). Sudden and forced outages, as unscheduled shutdowns of operations, are often consequences of equipment failure. Again, to bring scales into line we arbitrarily choose N=200, K=10.

[1] It is pointed out that this is sanitized data for illustration purposes only. It in no way reflects the actual operating positions of the various plants.

Table 17-2. Outputs and Inputs for Unit Level Analyses

Group	Unit	Outputs			Inputs	
		OPER	OUT	EQDER	MAINT	OCCUP
Plant 1	1	573	95	110	538	895
	2	560	138	120	290	770
	3	637	151	150	386	886
	4	685	139	160	290	760
	5	542	157	130	343	721
	6	520	100	120	470	810
	7	531	122	60	439	820
	8	511	135	160	293	888
Plant 2	1	521	102	93	440	771
	2	634	93	102	324	780
	3	610	86	75	378	825
	4	538	95	106	380	815
	5	591	116	119	241	880
	6	650	123	105	141	766
	7	621	107	91	355	823
	8	686	125	110	270	750
Plant 3	1	620	120	130	350	750
	2	550	81	95	630	770
	3	525	105	125	495	860
	4	580	125	106	345	800
Plant 4(1)	1	430	105	140	190	810
Plant 4(2)	1	560	110	105	280	770
	2	510	125	95	180	820
Plant 5	1	650	170	140	300	7000
	2	550	120	120	275	800
	3	580	160	110	447	650
	4	640	110	130	370	720
Plant 6	1	480	95	125	228	880

Table 17-2 continued

Group	Unit	Outputs			Inputs	
		OPER	OUT	EQDER	MAINT	OCCUP
Plant 7(1)	1	320	70	110	230	790
	2	250	60	110	220	790
	3	370	100	140	320	840
	4	280	90	100	280	810
Plant 7(2)	1	520	120	100	281	750
	2	430	100	140	302	850
	3	470	110	150	227	770
	4	410	80	110	254	825
Plant 8	1	475	100	120	179	750
	2	560	150	120	143	800
	3	510	120	110	114	750
	4	425	140	90	172	820

- EQDER - a function of forced deratings caused by equipment failure.

EQDER = N-K (# equipment related deratings), with N=200 and K=10 as above.

Since on the output side, any measure used must be such that bigger is better, one cannot *directly* take outages as an output. To achieve the bigger is better condition, we subtract outages from some constant to create a proper scale measure. The value 200 has been chosen arbitrarily, but at the same time to yield "OUT" values that are in line with the scales used for other factors. Some sensitivity analyses were done relative to this parameter (200), and the particular value chosen was found to have very little effect on the final relative efficiency outcomes.

Inputs

- MAINT - the total maintenance expenditure (labor + materials) in thousands of dollars.

Clearly, we could separate this into monetary inputs, but for purposes here we aggregate the two amounts into one figure.

- OCCUP - a function of total occupied hours, that is

OCCUP = $\frac{1}{10}$ (Total hours available - available but not operating hours).

In evaluating the ten level 2 DMUs (where, for example, the group of plants at Plant 1 is taken as a DMU), the averages of the level 1 DMUs make up the first three outputs and the first two inputs. For example, the average of the ten Plant 1 operating hours figures is 582. In addition to these aggregated figures, two further outputs, ENDER (a factor for environmental deratings) and planned capacity were used for the level 2 analyses. As well, a third input, average year of construction, was utilized. The data for the level 2 analyses is shown in Table 17-3.

Table 17-3. Group Level Data

	Outputs					Inputs		
Group	Oper.	Out.	Eqder	Ender	Cap	Maint.	Occup	Yr. const.
Plant 1	582	130	126	125	500	381	818	71
Plant 2	606	106	100	147	300	317	801	68
Plant 3	569	103	108	121	500	455	795	70
Plant 4(1)	430	105	140	111	100	190	810	65
Plant 4(2)	420	105	100	125	150	350	815	75
Plant 5	605	140	125	141	500	348	717	74
Plant 6	480	95	125	117	200	348	800	78
Plant 7(1)	305	80	115	110	100	263	808	56
Plant 7(2)	458	103	125	116	200	266	799	58
Plant 8	493	128	110	135	50	152	780	52

17.5.1 Hierarchical Analysis

Table 17-4 displays the outcomes from the hierarchical analysis. Here, power units have been grouped by location (Plant 1, Plant 2, ..., Plant 8), and have been analyzed using the hierarchical DEA model (17.2a)-(17.2g) and (17.2a$'$),(17.2b$'$)). The 10 group ratings are shown under column (3). Column (4) provides the "within group" ratings of individual power units, i.e., those ratings achieved when units are compared only to the members of their own group. To obtain ratings whereby all 40 DMUs can be compared on a common basis, the suggested three-stage adjustment developed earlier has been applied to the column 4 figures. The resulting adjusted values are shown in column 5.

Table 17-4. Efficiency Scores - Hierarchical Analysis (Grouped by Location)

(1) Group	(2) Unit	(3) Group Ratings	(4) Unit Ratings	(5) Adjusted Unit Ratings	(1) Group	(2) Unit	(3) Group Ratings	(4) Unit Ratings	(5) Adjusted Unit Ratings
Plant 1	1	100.0	70.8	70.8	Plant 4(1)	1	100.0	100.0	87.7
	2		99.1	99.1	Plant 4(2)	1	80.6	100.0	70.7
	3		86.0	86.0		2		100.0	70.7
	4		100.0	100.0	Plant 5	1	100.0	100.0	90.2
	5		100.0	100.0		2		93.5	84.3
	6		71.1	71.1		3		100.0	90.2
	7		76.1	76.1		4		95.7	86.3
	8		98.7	98.7	Plant 6	1	87.5	100.0	76.8
Plant 2	1	100.0	82.0	80.2	Plant 7(1)	1	87.1	100.0	76.4
	2		89.1	87.1		2		100.0	76.4
	3		80.7	78.9		3		100.0	76.4
	4		88.5	86.5		4		100.0	76.4
	5		95.1	93.0	Plant 7(2)	1	93.9	100.0	90.5
	6		100.0	97.8		2		84.5	76.4
	7		82.4	80.5		3		100.0	90.5
	8		100.0	97.8		4		79.6	72.0
Plant 3	1	100.0	100.0	94.9	Plant 8	1	100.0	100.0	89.7
	2		86.1	81.7		2		100.0	89.7
	3		83.7	79.4		3		100.0	89.7
	4		100.0	94.9		4		90.9	81.6

17.5.2 Hierarchical Analysis

In the above analyses, power units were grouped by location (e.g., the 8 Plant 1 units formed one group). The within groups analyses resulted in the ratings shown in column 4 of Table 17-4. Two other types of groupings were then evaluated - by fuel type and by capacity. Table 17-5 specifies the memberships of the groups. When the within group analyses were carried out on the power units under these alternative groupings, ratings of units changed to reflect group membership. Table 17-6 displays power unit ratings under the different membership scenarios (columns (2),(3),(4)). The location scenario has been replicated here (from Table 17-4). To combine the three ratings for each power unit, model (17.4a)-(17.4d) and model (17.4e) with (17.4b)-(17.4d) were applied. The outcomes from these models are displayed under columns (5) and (6) respectively. In both instances the set Φ of (17.4d) is defined such that capacity is rated to be of highest importance, followed by location, then by fuel type, i.e.,

$$\text{Capacity} > \text{location} > \text{fuel.}$$

Although multiple rankings of attributes could clearly be applied to this example, such an analysis was not carried out here.

Table 17-5. Plant Groupings by Capacity and Plant Groupings by Fuel Type

Group	Capacity	Units Included
Group 1	500 MWH	Plant 1, Plant 3, Plant 5
Group 2	200-300 MWH	Plant 2, Plant 6, Plant 7(2)
Group 3	< 200 MWH	Plant 4(1), Plant 4(2), Plant 7(1), Plant 8

Group	Fuel Type	Units Included
Group 1	U.S. Bit. Coal	Plant 1, Plant 2, Plant 3, Plant 4(1), Plant 8
Group 2	Gas/Coal	Plant 7(1), Plant 7(2)
Group 3	Liquid Bit. Coal	Plant 4(2)
Group 4	Oil	Plant 5
Group 5	Liqnite Bit. Coal	Plant 6

Table 17-6. Power Unit Ratings Under Different Groupings

(1)		(2)	(3)	(4)	(5)	(6)
		Grouping by				
Plant	Unit	Location	Capacity	Fuel	Aggregate (District Wts.)	Aggregate (Common Wts.)
1	1	70.8	68.8	69.6	69.8	69.6
	2	99.1	87.1	90.7	93.1	91.7
	3	86.0	82.0	86.0	84.7	84.0
	4	100.0	100.0	100.0	100.0	100.0
	5	100.0	90.0	100.0	96.7	95.0
	6	71.1	70.8	71.0	71.0	71.0
	7	76.1	69.6	76.1	73.9	72.8
	8	99.5	98.7	99.5	99.0	98.8
2	1	82.0	80.1	73.6	81.0	79.6
	2	89.1	88.9	87.4	89.0	88.7
	3	80.7	80.7	78.1	80.7	80.3
	4	88.5	80.4	72.1	84.4	81.7
	5	95.1	85.0	78.4	90.0	87.3
	6	100.0	100.0	100.0	100.0	100.0
	7	82.4	82.4	80.6	82.4	82.1
	8	100.0	100.0	100.0	100.0	100.0
3	1	100.0	89.0	90.7	94.5	92.9

	2	86.1	76.6	76.9	81.3	79.8
	3	83.7	68.9	68.9	76.3	73.8
	4	100.0	78.0	82.7	89.0	86.1
4(1)	1	100.0	100.0	100.0	100.0	100.0
4(2)	1	100.0	100.0	100.0	100.0	100.0
	2	100.0	88.1	100.0	96.0	94.0
5	1	100.0	100.0	100.0	100.0	100.0
	2	93.5	86.6	93.5	91.2	90.0
	3	100.0	100.0	100.0	100.0	100.0
	4	95.7	95.6	95.7	95.6	95.6
6	1	100.0	80.6	100.0	93.5	90.3

(1)		(2)	(3)	(4)	(5)	(6)
Plant	Unit	Grouping by Location	Capacity	Fuel	Aggregate (District Wts.)	Aggregate (Common Wts.)
7(1)	1	100.0	80.5	72.3	90.2	85.6
	2	100.0	80.4	75.5	90.2	86.1
	3	100.0	96.2	85.4	98.1	95.7
	4	100.0	74.5	73.7	87.2	82.9
7(2)	1	100.0	95.8	100.0	98.6	97.9
	2	84.5	84.5	84.5	84.5	84.5
	3	100.0	100.0	100.0	100.0	100.0
	4	79.6	72.8	79.6	77.3	76.2
8	1	100.0	100.0	93.0	100.0	98.8
	2	100.0	100.0	100.0	100.0	100.0
	3	100.0	100.0	100.0	100.0	100.0
	4	90.9	90.9	89.4	90.9	90.7

17.6. SIMULTANEOUS EVALUATION ACROSS LEVELS

The model discussed above evaluates efficiencies at various levels in a hierarchy in a multi-stage fashion. Specifically, in stage 1, performance measures for power units within each plant are computed relative to their peers (within that plant's subset of units). In stage 2, the plants, at Level 2, are treated as DMUs, and requisite efficiency scores are computed there.

Level 1 scores (for the power units) are then adjusted to reflect differences in efficiencies among the plants. In the hierarchical structure, DMUs at Level n have 2 types of inputs and outputs: (1) those consisting of aggregates of the corresponding factors at Level $n-1$, and (2) additional measures that apply only at Level n.

In the current section we approach efficiency measurement at the various levels in this hierarchical structure by considering all levels simultaneously, and by directing the optimization at the highest level in the hierarchy. In the two-level setting, this means treating the plants at Level 2 as the DMUs, with the power units at Level 1 viewed as *components* of the DMUs. The complicating feature of this approach is the presence of plant-specific output factors which must be apportioned across the components in an equitable manner. The ideas used here are similar to those applied in Chapter 6 involving multi component efficiency in banking.

There appear to be at least two disadvantages of the two-stage approach discussed above. First, the measure applied (as suggested by the power authority) is simply related to the *frequency* of environmental deratings per year, as opposed to some function of the level of the SO_2 above or below the threshold. Arguably, it is the *quantity* of environmental damage that one may wish to capture as an output from the plant. Second, since the environmental variable only applies at the plant level, it is then the case in the hierarchical model that each power unit within that plant is *equally* penalized. Clearly, however, an individual power unit in a plant may contribute more or less toward the production of hazardous materials (e.g. SO_2) than is true for some other power unit. A power unit that is, for example, shut down for maintenance during peak pollution periods would not likely contribute as much to pollution accumulation as other units that were operating at full capacity during that time.

In this section, we present an augmented version of the DEA model that views both levels in the hierarchy simultaneously, generating performance measures for each plant and for the power units within those plants. Level 2 (plant level) variables are allocated across the level 1 power units. This is done in a manner consistent with any imposed constraints on the proportions of the output assigned to the various power units, and with the objective of maximizing the performance measure of the level 2(plant) unit under consideration at any stage in the DEA model.

Consider the situation in which there are K power plants, with J_k power units within plant k. We define:

$Y_{kj_k} = (y_{kr_1 j_k})$ - the R_1 - dimensional vector of outputs generated by power unit j_k in plant k.

$X_{kj_k} = (x_{kij_k})$ - the I - dimensional vector of inputs consumed by power unit j_k in plant k.

$Y_{ks} = (y_{kr_2s})$ - the R_2 - dimensional vector of outputs generated by plant k.

Let ν, μ, μ_s denote vectors of multipliers associated with X_{kj_k}, Y_{kj_k} and Y_{ks} respectively.

It is noted that in the current application, plant level (level 2) factors appear only on the output side. In the previous model, year of construction was taken as a level 2 input, but turned out to be relatively insignificant. While the model structure herein is easily extended to include both inputs and outputs, we restrict our attention only to such factors on the output side.

To facilitate model development, define the R_2-dimentional decision vectors $\alpha^k_{j_k} = (\alpha^k_{r_2 j_k})$, where $\alpha^k_{r_2 j_k}$ is the proportion of output y_{kr_2s} allocated to power unit j_k. As well, let Y_k, X_k denote the aggregates of the output vectors $\{Y_{kj_k}\}_{j_k}$ and input vectors $\{X_{kj_k}\}_{j_k}$, respectively. That is

$$Y_k = \sum_{j_k \varepsilon J_k} Y_{kj_k}, X_k = \sum_{j_k \varepsilon J_k} X_{kj_k}.$$

In this particular problem setting, aggregates derived in this manner make logical sense, although in some settings, sums of outputs may not be relevant.

The proportion $\alpha^k_{r_2 j_k}$ of output y_{kr_2s} to be allocated to power unit j_k, may fall within certain logical bounds. Arguably, in the case that a given output r_2, is, for example, SO_2 emissions, the relative shares of this output allocated to two given units j_{k_1}, j_{k_2} could depend on a number of factors. These would include fuel types used, capacities in megawatt hours, operating hours, frequency of equipment failure deratings, etc. Since fuel type and capacity are fixed for units within the same plant, one can assume that $\alpha^k_{r_2 j_k}$ is a function of factors such as operating hours. Reasonable bounds might take the form:

$$L_{j_k} \le \alpha^k_{r_2 j_k} / \alpha^k_{r_2 1} \le U_{j_k}$$

Here, we assume that power unit #1 in plant k is taken as a standard, and other units j_k are compared to #1. L_{j_k} and U_{j_k} represent lower and upper limits respectively on the ratio of the proportions of output r_2 assigned to power units #1 and # j_k.

In the present two-level structure as described earlier, the plant (k) level performance measure (for any given set of multipliers (μ, μ_s, ν)) is given by:

$$e^k = [\mu Y_k + \mu_s Y_{ks}] / \nu X_k \qquad (17.7)$$

Here, we distinguish between Y_k, the aggregate of level 1 (power unit) outputs, and Y_{ks}, the plant level (level 2) outputs that are to be allocated to the respective level 1 units. We can view Y_{ks} as a form of *shared* output (that is, shared among the power units). The *corresponding* j_k^{th} power unit performance ratio is given by:

$$e^k_{j_k} = [\mu Y_{kj_k} + \mu_s \alpha^k_{j_k} Y_{ks}] / \nu X_{kj_k} \quad (17.8)$$

We use here the notation $\alpha^k_{j_k} Y^k_s$ to denote the R_2-dimensional vector $[\alpha^k_{1j_k} y_{k1s}, \alpha^k_{2j_k} y_{k2s}, \ldots, \alpha^k_{R_2 j_k} y_{kR_2 s}]^t$.

Property 17.3: The aggregate performance measure e^k of (17.7) is a convex combination of the J_k power unit measures $\{e^k_{j_k}\}^{J_k}_{j_k=1}$, defined in (17.8).

Property 17.4: A power plant k is efficient ($e^k = 1$) if and only if each power unit j_k within the plant is efficient ($e^k_{j_k} = 1$).

We now propose the following two-level variant of the standard CCR model:

$$\begin{aligned}
&\max e^o \\
&\text{subject to:} \\
&e^k \le 1 && \text{all } k, \\
&e^k_{jk} \le 1 && \text{all } k, j_k \,\varepsilon\, J_k \\
&L_{jk} \le \alpha^k_{r_2 j_k} / \alpha^k_{r_2 1} \le U_{jk} && \text{all } r_2, k, j_k \varepsilon J_k, \\
&\sum_{j_k \varepsilon J_k} \alpha^k_{r_2 j_k} = 1 && \text{all } r_2, k, \\
&\alpha_{r_2 j_k}, \mu_{r_1}, \mu_{sr_2}, v_i \ge 0, && \text{all } r_1, r_2, k, j_k i.
\end{aligned} \quad (17.9)$$

Problem (17.9) is nonlinear in two respects. First e^k and $e^k_{j_k}$ are linear fractional functionals. Second $e^k_{j_k}$ involves the product of variables $\mu_{sr_2} \alpha^k_{r_2 j_k}$. However, it can be shown that (17.9) is equivalent to a linear programming formulation, as given by the following theorem.

Theorem 17.2:
Problem (17.9) can be represented as a linear programming problem.

Proof:
First it is noted that from Property (17.4), the constraints $e^k \le 1$ are redundant, and can be removed from the problem. Make the change of variables

$$\gamma_{kr_2 j_k} = \mu_{sr_2} \cdot \alpha^k_{r_2 j_k}.$$

It is noted that the constraint set

$$L_{j_k} \le \alpha^k_{r_2 j_k} / \alpha^k_{r_2 1} \le U_{j_k}$$

becomes

$$L_{j_k}\alpha^k_{r_2 1} \le \alpha^k_{r_2 j_k} \le U_{j_k}\alpha^k_{r_2 1}$$

which, with multiplication through by μ_{sr_2}, becomes

$$L_{jk}\gamma_{r_2 1} \le \gamma_{kr_2 j_k} \le U_{j_k}\gamma_{r_2 1}.$$

As well, the convexity restriction

$$\sum_{j_k \varepsilon J_k} \alpha^k_{r_2 j_k} = 1$$

can be replaced by

$$\sum_{j_k \varepsilon J_k} \gamma^k_{r_2 j_k} = \mu_{sr_2}.$$

Following the standard conversion of Charnes and Cooper (1962) the linear fractional programming model (17.9) becomes

$$\begin{aligned}
&\max\ \mu Y^o + \mu_s Y^o_s,\\
&\text{subject to:}\\
&vX_o = 1,\\
&\mu Y_{kj_k} + \gamma^k_{j_k} Y_{ks} - vX_{kj_k} \le 0 \quad \text{all } k, j_k,\\
&L_{j_k}\gamma^k_{r_2 1} \le \gamma^k_{r_2 j_k} \le U_{j_k}\gamma^k_{r_2 1} \quad \text{all } r_2, k, j_k\\
&\gamma^k_{r_2 j_k}, \mu_r, \mu_{sr_2}, v_i \ge 0 \quad \text{all } r_1, r_2, k, j_k, i.
\end{aligned} \tag{17.10}$$

Clearly, problem (17.10) satisfies the necessary linearity property.

QED

From the optimal solution of (17.10), one can compute $\hat{\alpha}^k_{r_2 j_k}$ from

$$\hat{\alpha}^k_{r_2 j_k} = \hat{\gamma}^k_{r_2 j_k} / \hat{\mu}_{sr_2}$$

In the following section, we apply model (17.10) to evaluate efficiencies of a set of power plants and corresponding power units.

17.7. ANALYSIS OF EFFICIENCY: AN EXAMPLE

Considering again the data of Table 17-2, we can view the power plants (level 2 in the hierarchy) as aggregates of the units that comprise those plants. In this regard, the aggregates of all level 2 outputs and inputs can serve as level 2 factors. (As discussed previously, such aggregation may not be relevant in all cases, although it is so in this instance). In addition, there are factors that pertain primarily to the plant level only. The best example of such a factor in this situation is SO_2 emissions. The total environmental damage caused by a plant can be measured by the level (density of particulates) of SO_2 above some tolerable threshold, and multiplied by the number of hours that this phenomenon prevails during the year. Again, this factor falls into the more is worse category, as is true of the level 1 outputs, and was subtracted from the worst case value.

17.7.1 Proportional split of plant-level outputs

In the process of solving (17.10), the γ_{j_k} - variables (that give rise to the α_{j_k} - variables) are intended to split the shared output (SO_2) across the units in a plant, in a way that is most fair for that plant. If a particular power unit j_{k_1} in a plant is experiencing a higher degree of outages and equipment-related deratings than is true of the other units in that plant, then j_{k_1} should arguably be penalized with a smaller proportion of the environmental damage due to SO_2.

Unfortunately, the data is too coarse to be able to detect when a power unit was simultaneously experiencing equipment-related deratings, and environmental (SO_2) deratings. Clearly, if a power unit j_{k_1} was shut down for some reason on a given day when SO_2 emissions were high, the corresponding α_{jk_1} should be set to 0. To capture this idea we have imposed assurance region constraints, as per Thompson et al. (1992), of the form:

$$L_{j_k} \leq \alpha^k_{r_2 j_k} / \alpha^k_{r_2 1} \leq U_{j_k},$$

where we have numbered that power unit 1 as the unit whose total OUT + EQDER is lowest. (This is the power unit whose total number of hours of outage + equipment deratings is highest). The argument is that for plant $k, \alpha^k_{r_2 1}$ should be the lowest proportion among all units for that plant. We have then chosen $L_{j_k} = 1$ for all units j_k. Since it is unclear what the precise relationship is concerning the timing of non-environmental deratings and outages (as discussed above), we have chosen here to set all U_{j_k} equal to one another. We experimented with different values, and found that while the efficiency ratings of the various power units within a plant tended to decrease as U_{j_k} is lowered, their order (relative to one another) was quite stable. Table 17-7 displays the plant-level and associated power unit-level efficiency scores.

The advantage of viewing efficiency in this manner is that not only can one evaluate the performance of plants, but at the same time can uncover the extent to which each of the subunits (power units) within the plant is contributing to that performance. This permits management to identify which power units in a plant are under-performing, and which units could serve as benchmarks within that plant. Following Properties 17.3 and 17.4, it is noted that for efficient plants such as 4,5 and 6, all power units within these are efficient as well.

Table 17-7. Power Plant & Power Unit Ratings

Plant	Unit	Unit Rating	Plant Rating	Plant	Unit	Unit Rating	Plant Rating
1	1	.64	.833	4(1)	1	1	1
	2	1		4(2)	1	1	1
	3	.99			2	1	
	4	1					
	5	1		5	1	1	1
	6	.67			2	1	
	7	.46			3	1	
	8	1			4	1	
2	1	.58	.861	6	1	1	1
	2	.80					
	3	.63		7(1)	1	.87	.84
	4	1		2	1		
	5	1		3	.81		
	6	1		4	.69		
	7	1					
	8	1		7(2)	1	.82	.935
					2	.92	
3	1	.82	.793		3	1	
	2	.91			4	1	
	3	.79		8	1	.99	.997
	4	.66			2	1	
					3	1	
					4	1	

17.8. CONCLUSIONS

This chapter has presented DEA-based models for evaluating the efficiency of a set of power plants, and corresponding power units as a hierarchical structure. In the earlier part of the chapter, hierarchical efficiency was viewed as a multi-stage process. In Section 17.7 however, hierearchical efficiency measurement is viewed at all levels simultaneously. This is accomplished by first defining the decision making units (DMUs), as the units at the highest level in the hierarchy (power plants in the current application). The elements lower down in the hierarchy are then viewed as components of the top level DMUs, and as such, have their efficiency evaluated as well.

A complicating feature of this latter structure is the presence of outputs at any level in the hierarchy that must be allocated among the components at the next stage down in that hierarchy. In the setting herein, this is accomplished by defining variables which provide for a split of such (plant-level) outputs among the power units within each plant. We demonstrate that this resulting non-linear model can be converted to a linear programming problem.

The developed models have been applied to 40 power generating units organized under 8 plants. Sulphur dioxide (SO_2) emissions are generally regarded as a plant-level output which we wish to allocate to the power units under each plant. This allocation in practice could be a function of various factors including the percent downtime for scheduled maintenance, etc. The outcome of the efficiency evaluation is given in Table 17-7.

The application of DEA principles to hierarchical structures is an important area for research. Many organizational structures tend to exhibit such a profile. The ideas herein can potentially lend themselves to other areas of study, for example, supply chains. The ideas are also somewhat related, as well, to the concepts presented by Fare and Grosskopf (1996) regarding intermediate products, as well as structures studied in the network DEA model of Fare and Grosskopf (2000).

REFERENCES

1. Ali, I., W. Cook, and L. Seiford (1991), Strict vs weak ordinal relations for multipliers in data envelopment analysis, *Management Science* 37(6): 733-738.
2. Banker, R., and R. Morey (1986), The use of categorical variables in data envelopment analysis, *Management Science* 32(12):1613-1627.
3. Charnes, A., and W.W. Cooper (1962), Programming with linear fractional functions, *Naval Research Logistics Quarterly* 9, 181-186.
4. Charnes, A., W.W. Cooper, and E.L. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2(6), 429-444.
5. Cook, W.D., D. Chai, J. Doyle, and R. Green (1998), Hierarchies and groups in DEA, *Journal of Productivity Analysis* 10, 177-198.
6. Cook, W.D. and R. Green (2004), Evaluating power plant efficiency: A hierarchical model, *Computers and Operations Research,* 32 (4), 813-823.
7. Cook, W.D., M. Kress, and L. Seiford (1996), Data envelopment analysis in the presence of both quantitative and qualitative factors, *Journal of the Operational Research Society* 47, 945-953.
8. Färe, R., and S. Grosskopf (1996), Productivity and intermediate products: a frontier approach, *Economic Letters* 50(1), 65-70.

9. Färe, R., and S. Grosskopf (2000), Network DEA, *Socio-Economic Planning Sciences* 34, 35-49.
10. Golany, B. (1988), A note on including ordinal relations among multipliers in data envelopment analysis, *Management Science* 34(8), 1029-1033.

This chapter is based upon Cook, W.D., D. Chai, J. Doyle, and R. Green, 1998, Hierarchies and groups in DEA, *Journal of Productivity Analysis*, 10:177-198, with permission from Kluwer Academic Publishers, and Cook, W.D. and R. Green, 2003, Evaluating power plant efficiency: A hierarchical model, *Computers and Operations Research*, 32 (4), 813-823, with permission from Elsevier Science.

SUBJECT INDEX

A

B

C

D

E

F

G

H

I

K

L

M

N

V

W

LICENSING AND WARRANTY AGREEMENT

READ THIS: Do not install or use the CD-ROM until you have read and agreed to this agreement. By opening the accompanying software packet, you acknowledge that you have read and accepted the following terms and conditions. If you do not agree and do not want to be bounded by such terms and conditions, do not install or use the software "DEAFrontier".

License: The copyright to the software in the CD-ROM (the "DEAFrontier") is owned by Joe Zhu. The "DEAFrontier" is protected by the United States copyright laws and international treaty provisions. No part of the DEAFrontier may be reproduced, stored in a retrieval systems, distributed (including but not limited to over the Internet), modified, decompiled, reverse engineered, reconfigured, transmitted, or transcribed, in any form or by any means without the permission of the author. The DEAFrontier may not, under any circumstances, be reproduced for sale. This license allows you to use the DEAFrontier for educational and research purposes only, not for commercial purposes. You may only (i) make one copy of the DEAFrontier for backup or archival purposes, or (ii) transfer the DEAFrontier to a single hard disk, provided that you keep the original for backup or archival purposes. You may not (i) rent or lease the DEAFrontier, (ii) copy or reproduce the DEAFrontier through a LAN or other network system or through any computer subscriber system, or (iii) modify, adapt, or create derivative works based upon the DEAFrontier. You may be held legally responsible for any copying or copyright infringement which is caused by your failure to abide by the above terms and conditions.

Limited Warranty: The warranty for the media on which the DEAFrontier is provided is for thirty (30) days from the original purchase and valid only if the packaging for the DEAFrontier was purchased unopened. If the Publisher receives notification within the warranty period of defects in materials or workmanship, the Publisher will replace the defective software media. The Publisher and the author provide no other warranties, expressed or implied, including without limitation implied warranties of merchantability and fitness for a particular purpose, and shall not be liable for any damages, including direct, special, indirect, incidental, consequential, or otherwise. The Publisher does not warrant that the functions contained in the DEAFrontier will meet your requirements or that the operation of the DEAFrontier will be error-free.

DEAFrontier Installation Instructions

The CD-ROM contains the DEAFrontier* which is a DEA Add-In for Microsoft Excel. This software "DEAFrontier" requires Excel 97 or later versions. Please read Chapter 1 for installation instructions.

*May not work on a Macintosh