International Series in
Operations Research & Management Science

William W. Cooper
Lawrence M. Seiford
Joe Zhu *Editors*

Handbook on Data Envelopment Analysis

Second Edition

Springer

International Series in Operations Research & Management Science

Volume 164

For further volumes:
http://www.springer.com/series/6161

William W. Cooper • Lawrence M. Seiford
Joe Zhu
Editors

Handbook on Data Envelopment Analysis

Second Edition

Springer

Editors
William W. Cooper
Department of Management Science
and Information Systems
McCombs School of Business
University of Texas at Austin
Austin, TX, USA
cooperw@mail.utexas.edu

Lawrence M. Seiford
Department of Industrial
and Operations Engineering
University of Michigan
Ann Arbor, MI, USA
seiford@umich.edu

Joe Zhu
School of Business
Worcester Polytechnic Institute
Worcester, MA, USA
jzhu@wpi.edu

Please note that additional material for this book can be downloaded from
http://extras.springer.com

ISSN 0884-8289
ISBN 978-1-4419-6150-1 e-ISBN 978-1-4419-6151-8
DOI 10.1007/978-1-4419-6151-8
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011934489

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Ruth
WWC

To Alec & Andie
JZ

Preface

This new edition updates previous editions of the *Handbook on Data Envelopment Analysis*. As noted in preceding editions, data envelopment analysis (DEA) is a "data-oriented" approach for evaluating the performances of a set of entities called decision-making units (DMUs) which convert multiple inputs into multiple outputs. As will be seen from the chapters in this (and the preceding) editions, DEA has been used in evaluating the performances of many different kinds of entities engaged in many different kinds of activities in many different contexts. It has opened up possibilities for use in cases which have been resistant to other approaches because of the complex and often unknown nature of the relations between the multiple inputs and outputs involved in many of their activities (which are often reported in noncommensurable units). See Emrouznejad et al. (2008) who "have identified more than 4,000 research articles published in journals or book chapters. . .. To provide some sense of the field's ongoing expansion, has had we included unpublished dissertations, working/research manuscripts, the bibliography count would have exceeded 7,000 entries!" All this occurred in the 30 years since the publication of the originating article by Charnes et al. (1978) and the pace continues to accelerate. DEA has also been used to supply new insights into activities and entities that have previously been evaluated by other methods.

This handbook is intended to represent another milestone in the progression of DEA. Written by experts, who are often major contributors to the topics to be covered, it includes a comprehensive review and discussion of basic DEA models, extensions to the basic DEA methods, and a collection of DEA applications in many different areas such as banking, service industries, health care and education as well as evaluations of country and regional performances, and engineering and science applications.

Handbook chapters are organized into two categories (1) basic DEA models, concepts, and their extensions and (2) DEA applications. The first category consists of 12 chapters. Chapter 1, by Cooper, Seiford and Zhu, covers the various models and methods for treating "technical" and "allocative" efficiency. It also includes the "additive" model for treating "allocative" and "overall" efficiency that can be used when the usual "ratio" form of the efficiency measure gives unsatisfactory

or misleading results. Chapter 2, by Banker and Cooper, deals with returns to scale (RTS) and the ways in which this topic is treated with different models and methods. The emphasis in this chapter is on relationships between models and methods and the RTS characterizations that they produce. This chapter also introduces a method for determining "exact" elasticities of scale in place of previous approaches, which are limited because they can only establish "bounds" on the elasticities. Chapter 3, by Cooper, Li, Seiford and Zhu, describes ways to determine the "sensitivity" and "stability" of DEA efficiency evaluations in the presence of stipulated variations in the data. The sensitivity analyses covered in this chapter extend from variations in *one* "data point" and include determining the sensitivity of DEA efficiency evaluations when *all* data points are varied simultaneously.

In Chap. 4, Cooper, Ruiz, and Sirvent describe different approaches with no need for a priori choices of weights (called "multipliers) that reflect meaningful trade-offs. It also shows how to incorporate prices or other value information and managerial goals, making a choice among alternate optima for the weights, avoiding the need for zero weights and avoiding large differences in the values of multipliers, improving discrimination and rankings of performances. Chapter 5, by Färe, Grosskopf, and Margaritis, provides an overview of recent work on DEA and Malmquist productivity indexes. It also reviews the construction of static and dynamic DEA technologies. Based on these technologies it shows how DEA can be used to estimate the Malmquist productivity index. Chapter 6, by Cook, discusses how to treat qualitative data in DEA. The emphasis is on cases in which the data are ordinal and not cardinal. This extends DEA so that it can treat problems in which the data can be ordered but the numbers utilized to represent the ordering do not otherwise lend themselves to the usual arithmetic operations such as addition, multiplication, etc.

Chapter 7, by Cooper, Deng, Seiford, and Zhu, treats "congestion" and discusses modeling to identify the amounts and sources of this particularly severe form of "technical" inefficiency. Chapter 8, by Tone, introduces a slacks-based model and its extensions. This is followed by three chapters directed to probabilistic and statistical characterizations of the efficiency evaluation models discussed in Chap. 1. Chapter 9, by Cooper, Huang, and Li, turns to probabilistic formulations as in "chance-constrained programming." It also treats "joint" chance constraints, as well as the more customary types. All of this is accompanied by discussions of uses of both types of constraints in some of the applications where these chance-constrained programming formulations of DEA have been used. Chapter 10, by Simar and Wilson, utilizes "bootstrapping" and shows how these methods may be used to obtain statistical tests and estimates of DEA results. Chapter 11, by Banker and Natarajan, is directed to the more classical methods of "statistically consistent estimates." Hence, both classical and more recently developed approaches are brought to bear on statistical characterizations that are now available for use with DEA.

The final chapter in the first category is written by Cook, Liang, and Zhu. An important area of development in recent years in DEA has been applications wherein internal structures of DMUs are considered. For example, DMUs may

consist of subunits or represent two-stage processes. One particular subset of such processes is those in which all the outputs from the first stage are the only inputs to the second stage. Chapter 12 reviews these models and discusses relations among various approaches. The focus here is the approaches based upon either Stackelberg (leader–follower) or cooperative game concepts.

The second category of the topics covered in this handbook involves four DEA applications. Chapter 13, by Paradi, Yang, and Zhu, provides a detailed discussion of DEA applications to banking with an emphasis on factors, circumstance, and formulations that need to be considered in actual applications. It also includes a comprehensive list of DEA bank branch models in the literature. Chapter 14, by Triantis, discusses DEA applications in engineering and includes a comprehensive bibliography of published DEA engineering applications. As this chapter shows, engineering uses of DEA have been relatively few but this is a field that is rich with potential applications of DEA ranging from engineering designs to uses of DEA to evaluate performances and to locate deficiencies in already functioning systems.

In contrast, the service sector holds substantial challenges for productivity analysis because most service delivery is often heterogeneous, simultaneous, intangible, and perishable. Chapter 15, by Avkiran, provides a selection of DEA applications in the service sector with a focus on building a conceptual framework, research design, and interpreting results. Chapter 16, by Chilingerian and Sherman, offers a succinct history of DEA to the provision of health care by, e.g., hospitals, and discusses the models and the motivations behind the applications with an eight-step application procedure and some "do's and don'ts" in DEA healthcare applications with an emphasis on the need for including "quality" measures of the services provided.

We hope this DEA handbook will serve as a comprehensive reference for researchers and practitioners and as a guide for further developments and uses of DEA. We welcome comments, criticisms, and suggestions.

Austin, TX — William W. Cooper
Ann Arbor, MI — Lawrence M. Seiford
Worcester, MA — Joe Zhu

About the Authors

Necmi K. Avkiran is an Associate Professor in financial studies at the UQ Business School, The University of Queensland, Brisbane, Australia. He received his MBA (Finance concentration) from Boston College. Following a year's employment in the banking sector he continued to undertake his PhD in banking from Victoria University, Melbourne. The main focus of his research has been bank performance measurement. He specializes in the applications of the productivity measurement technique DEA in the service sector. He holds an award for excellence from the MCB University Press for a research paper on bank customer service quality, and his publication on bank networks has been identified by Thomson Reuters Essential Science Indicators as a featured Fast-Breaking Paper in the field of Economics & Business, August 2010. In May 2006, he published the third edition of the research book entitled *Productivity Analysis in the Service Sector with Data Envelopment Analysis*. The journals in which he has published include the *Journal of Banking and Finance*, *Journal of Economics and Finance*, *Pacific-Basin Finance Journal*, *OMEGA*, *Socio-Economic Planning Sciences*, *Scientometrics*, *International Journal of Bank Marketing* (UK), *Journal of Asia-Pacific Business*, *International Journal of Human Resource Management*, *Personnel Review*, and *Asia Pacific Journal of Human Resources*, as well as on-campus publications. He is a member of the Australian Society for Operations Research, and the Financial Services Institute of Australasia (Australian Institute of Banking and Finance).

Dr. Rajiv D. Banker is the Merves Chair in Accounting and Information Technology at the Fox School of Business, Temple University. Dr. Banker is internationally recognized as one of the most prolific and influential leaders in interdisciplinary research in management. His research addresses complex and emerging problems of importance to managers. He is a world-renowned scholar, an innovative leader, and a distinguished teacher. Dr. Banker is one of the most highly cited scholars in management and economics worldwide. He is recognized by the Institute for Scientific Information (Web of Science) as one of the most influential researchers in economics and business worldwide (comprising less than one-tenth of 1% of all publishing researchers in the past three decades) that have

made fundamental contributions to the advancement of science and technology. His research articles are cited over 200 times each year by other researchers in a wide range of disciplines. One of his papers is ranked fourth highest in citations in the 50-year history of Management Science. He is the President of the International Data Envelopment Analysis Society and the editor-in-chief of the *Journal of DEA*.

Jon A. Chilingerian is a tenured Professor at the Heller School at Brandeis University. He began working with DEA as a Doctoral Student at Massachusetts Institute of Technology (MIT) while working as a research and teaching assistant for Dr. David Sherman. He is director of the MD-MBA Program and Director of the AHRQ-funded Doctoral Training Program in Health Services Research at Brandeis University. He is a visiting Professor of Health Care Management and member of the Health Management Initiative at INSEAD in Fontainebleau, France. He teaches graduate courses and executive education sessions in Organizational Theory and Behavior, Management of Health Care Organizations, and Health Services Research. Dr. Chilingerian is the co-author of International Health Care Management, published by Elsevier Press (Summer 2005), and The Lessons and the Legacy of the Pew Health Policy Program, with Corinne Kay, published in 1997 by the Institute of Medicine National Academy Press. He has scholarly papers and review essays published in journals such as *Annals of Operational Research*, *Medical Care*, *European Journal of Operational Research*, *Health Services Research*, *Health Care Management Review*, *Medical Care Research and Review*, *Inquiry*, *Health Services Management Research*, and *The Journal of Health Politics, Policy and Law*. Dr. Chilingerian was former chair of the Health Care Management Division of the Academy of Management. His research focuses on managing health care organizations, ranging from studies of executive leadership and management of professionals to the measurement of performance (i.e., productive efficiency, quality, etc.), identification of physician best practices, and the analysis of effective operating strategies. He is currently working on advancing clinical applications of DEA by studying quality, productivity, and technical change in a variety of procedures such as orthopedic, cardiac, and breast cancer surgeries.

Wade Cook is the Gordon Charlton Shaw Professor of Management Science in the Schulich School of Business, York University, Toronto, Canada, where he serves as Department Head of Management Science and as Associate Dean of Research. He holds a doctorate in Mathematics and Operations Research. He has published several books and more than 130 articles in a wide range of academic and professional journals, including *Management Science*, *Operations Research*, *JORS*, *EJOR*, *IIE Transactions*, etc. His areas of specialty include DEA, and multicriteria decision modeling. He is a former Editor of the *Journal of Productivity Analysis*, and of *INFOR*, and is currently an Associate Editor of *Operations Research*. Professor Cook has consulted widely with various companies and government agencies.

William W. Cooper is the Foster Parker Professor of Finance and Management (Emeritus) in the Red McCombs School of Business, University of Texas at Austin. Author or coauthor of 23 books and 490 scientific professional articles, he has written extensively on DEA and consulted on its uses with numerous business firms and government agencies. He holds honorary DSc degrees at Ohio State and Carnegie Mellon Universities in the USA and the degree of Doctor Honoris Causa from the University of Alicante in Spain. A fellow of the Econometric Society and of INFORMS (Institute of Operations Research and Management Science), he is also a member of the Accounting Hall of Fame.

Honghui Deng is an Associate Professor in the School of Business, University of Nevada Las Vegas. He received his PhD in Business Administration from the Red McCombs School of Business, the University of Texas at Austin. His research focuses mainly on Operations Research/Management, Economics, Information Systems, Management Science, and Risk Management with both methodology and empirical studies. Dr. Deng's research has appeared or will be forthcoming in the *Management Science*, *Decision Sciences*, *Manufacturing & Service Operations Management*, *European Journal of Operational Research*, *Journal of Productivity Analysis*, and *Socio-Economic Planning Sciences* and among others. Dr. Deng has served as a member of several professional organization committees such as the China Summer Workshop on Information Management (CSWIM). He is currently also working as a consultant for the Institute of Innovation, Creativity and Capital (IC2) of the University of Texas at Austin. In addition, Dr. Deng works as joint/part time professor in several top-ranked Chinese universities such as Guanghua School of Management, Peking University, and so on. He is also the Director of the Joint Institute of The Innovation & Commercialization between IC2 and the Chinese University of Hong Kong which is located in Shenzhen, China.

Rolf Färe is Professor of Economics and Agricultural Economics at Oregon State University. After receiving his Filosofie Licentiat from Lund University under Professor Bjorn Thalberg, he continued his studies at U.C. Berkeley under Professor Ronald W. Shephard. There his work with Shephard laid the foundations for an axiomatic approach to production theory. Over the years this has led to the publication of ten books and more than 200 articles. See also Who's Who in Economics.

Shawna Grosskopf is Professor of Economics at Oregon State University. She received her MS and PhD at Syracuse University where her areas of specialization were public economics and the economics of education. Over the last 20 years she has published books and articles in the area of performance measurement and its applications to her areas of interest. See also Who's Who in Economics and ISI Most Highly Cited.

Zhimin Huang is Professor of operations management in the School of Business at Adelphi University. He received his BS in Engineering from The Beijing University of Aeronautics and Astronautics, MS in Economics from The Renmin University of China, and PhD in Management Science from The University of Texas at Austin. His research interests are mainly in supply chain management, DEA, distribution

channels, game theory, chance-constrained programming theory, and multicriteria decision-making analysis. He has published articles in *Naval Research Logistics*, *Decision Sciences*, *Journal of Operational Research Society*, *European Journal of Operational Research*, *Journal of Economic Behavior and Organization, Optimization, OMEGA*, *Research in Marketing*, *Annals of Operations Research*, *International Journal of Systems Science*, *Journal of Productivity Analysis*, *Journal of Economics*, *Journal of Mathematical Analysis and Applications*, *Computers and Operations Research*, *International Journal of Production Economics*, and other journals.

Shanling Li is Professor of Operations Management/Management Science and the Associate Dean of Research and International Relations at Desautels Faculty of Management of McGill University, Canada. She currently serves as the Director of Management Science Research Center at the Faculty. She received her PhD in Operations Management (OM) with minor in Operations Research from University of Texas at Austin, and MS in Industrial Management from Georgia Institute of Technology. Li's research interests are mainly in three directions: (1) interface of OM with finance/marketing; (2) supply chain management under uncertainties; and (3) productivity analysis. Her works appeared in *Management Science* (MS), *Operations Research* (OR), *Manufacturing & Service Operations Management* (M&SOM), *Production and Operations Management* (POM), *Journal of Operations Management* (JOM), *Information Systems Research* (ISR), *European Journal of Operational Research* (EJOR) and others. Li sits on the editorial boards of several journals including POM.

Susan X. Li is Professor of Management Information Systems in the School of Business at Adelphi University. She received her BS in Economics from The Renmin University of China and her PhD in Management Science and Information Systems from the Graduate School of Business at The University of Texas at Austin. Her research interests are mainly in supply chain-related information systems, distribution channels, DEA, chance-constrained programming theory in investment and insurance portfolio analysis, and multicriteria decision-making analysis. She has published articles in *Decision Sciences*, *Journal of Operational Research Society*, *European Journal of Operational Research*, *Annals of Operations Research*, *OMEGA*, *Information Systems and Operational Research*, *Computers and Industrial Engineering Journal*, *Journal of Optimization Theory and Applications*, *Journal of Productivity Analysis*, *Computers and Operations Research*, *International Journal of Production Economics*, and other journals.

Liang Liang is Professor at School of Management at University of Science and Technology of China. His research interests include issues of Multicriteria Decision Making and Supply chain, and DEA. He has published over 50 articles in journals such as *Operations Research*, *IIE Transactions*, *IEEE Transactions on SMC*, *Naval Research Logistics*, *European Journal of Operational Research*, *Annals of Operations Research*, *OMEGA*, *Journal of Productivity Analysis*, and others. He was supported by National Science Foundation for Distinguished Young Scholars of China. He is currently engaged as Distinguished Professor of Chang Jiang Scholar Program by Ministry of Education of P. R. China in 2010.

Dimitri Margaritis is Professor of Finance at the University of Auckland Business School. After receiving his PhD in Economics from SUNY-Buffalo he taught at SIU-Carbondale, the University of British Columbia and the University of Washington before taking up a position as Head of Research and Senior Research Fellow at the Reserve Bank of New Zealand. His research areas include efficiency and productivity measurement, corporate finance, and financial econometrics.

Ram Natarajan is an Associate Professor of Accounting and Information Management at the University of Texas at Dallas. Dr. Natarajan received a Post Graduate Diploma in Management from the Indian Institute of Management, Calcutta, and a PhD in Accounting from the Wharton School of University of Pennsylvania. Dr. Natarajan's research addresses questions pertaining to analysis of productivity and efficiency in organizations, the determinants of CEO compensation, use of accounting performance measures for valuation, and performance evaluation and properties of earnings forecasts made by financial analysts. His academic honors include a Dean's Fellowship of Distinguished Merit from the Wharton School of University of Pennsylvania, a dissertation Fellowship from Ernst and Young, a runner-up award for the outstanding doctoral dissertation from the Management Accounting Section of the American Accounting Association, and a Junior Faculty Teaching and Research Fellowship from Arthur Andersen. He has presented his research in many US universities and academic conferences.

Joseph C. Paradi was educated at the University of Toronto, Canada, where he received his BASc, MASc, and PhD degrees in Chemical Engineering and he is a Professional Engineer. He spent 20 years in the computer services industry where he was a founder and President of a large Canadian firm. He is the director of several high-tech firms and involved in venture capital activities. Since 1989 he has been at the University of Toronto where he is a Professor and Executive Director of the Centre for Management of Technology and Entrepreneurship in Toronto, Canada. His research focus is on the measurement and improvement of productivity, efficiency, and effectiveness in the services elements of telecommunications and financial services firms using DEA. He has published papers in *EJOR*, *IEEE Engineering Management*, *OMEGA*, *Annals of OR*, *JPA*, and other peer reviewed Journals. He is a member of the IEEE, IAMOT, INFORMS, IIE, and several other academic-oriented organizations.

José L. Ruiz is Associate Professor of Statistics and Operations Research at the University Miguel of Elche, Spain. He holds a PhD from this university. He develops his research at the Universitary Institute for Research "Centro de Investigación Operativa" (CIO), the analysis of efficiency and productivity being his primary research interest. He has published his research in several journals including *Operations Research*, *European Journal of Operational Research*, *The Journal of the Operational Research Society*, *Journal of Productivity Analysis*, *Fuzzy Sets and Systems*, *International Journal of Information Technology and Decision Making*, and others.

Lawrence M. Seiford is Professor of Industrial and Operations Engineering and the Goff Smith Co-Director of the Tauber Institute for Global Operations at the University of Michigan. His research interests are primarily in the areas of quality engineering, productivity analysis, process improvement, multicriteria decision making, and performance measurement. He has written and co-authored four books and over 100 articles in the areas of quality, productivity, operations management, process improvement, decision analysis, and decision support systems. Professor Seiford is past Editor-in-Chief of *OMEGA*, the *International Journal of Management Science*, *Senior Editor of the Journal of Data Envelopment Analysis*, *Associate Editor of the Journal of Productivity Analysis*, and has been or is on the editorial boards of nine scientific journals. Dr. Seiford has received the General Electric Outstanding Teaching Award, the CBA Foundation Award for Research Excellence, and has been a Lily Endowment Teaching Fellow. He was awarded the degree Docteur Honoris Causa from the National Ministry of Education of France in a special recognition ceremony at the Universite de la Mediterranee, Aix-Marseille II in November, 2000. He is a Fellow of the Institute of Industrial Engineers (IIE), a Fellow of the American Society for Quality (ASQ), and a Fellow of the Institute for Operations Research and the Management Sciences (INFORMS).

H. David Sherman is a tenured Full Professor at Northeastern University, College of Business Administration (Boston, Massachusetts). In addition to work on health care, David managed several DEA applications in banks (three of the larges 50 US banks), brokerage firms, and government organizations. Four of these DEA applications generated substantial benefits documented in the press. David is co-author with Professor Joe Zhu of *Service Productivity Management: Improving Service Performance using Data Envelopment Analysis* (Springer, Boston 2006). He currently serves on the Board and chair of the audit committee of two public companies and has consulted with government, banks, and technology businesses. He is the author of the *Fair Value Measurement Answer Book* (CCH 2010) and is co-author of *Profits You Can Trust: Spotting and Surviving Accounting Landmines* (2003-Prentice Hall-Financial Times). His research has been published in academic and management journals including *Harvard Business Review* (four articles), *Sloan Management Review*, *Accounting Review*, *European Journal of Operations Research*, *Annals of Operations Research*, *Bankers Magazine*, *Interfaces*, *Management Accounting*, *Journal of Banking and Finance*, *American Banker*, *Medical Care, and Auditing*.

Léopold Simar is Emeritus Professor of Statistics at The Univeristé Catholique de Louvain, Louvain-la-Neuve, Belgium, where he has been for 12 years as the Founder-Chairman of the Institute of Statistics. He has also been Professor and Dean for 12 years at The Facultés Universitaires Saint-Louis, Brussels, Belgium. He was teaching mathematical statistics, multivariate analysis, bootstrap methods in statistics, and econometrics. His research focuses on nonparametric and semiparametric methods (boundary estimation, density estimation, single index models, bootstrap), and multivariate statistics. He wrote more than 100 papers published in

the best international journals. He is an elected member of the ISI (International Statistical Institute) and the past President of the Belgian Statistical Society from which he has been elected Honorary Member. He received a "Chiarra Fama" professorship in Italy (University of Pisa) and a "Research Chair of Excellency Pierre de Fermat" in France (Toulouse School of Economics). He is currently associate editor of the *Journal of Productivity Analysis*.

Inmaculada Sirvent is Associate Professor of Statistics and Operations Research at the University Miguel of Elche, Spain, and develops her research at the Universitary Institute for Research "Centro de Investigación Operativa" (CIO). She holds a PhD from this university. Her primary research interest is the analysis of efficiency and productivity, especially with DEA models. She has published her research in journals such as *Operations Research*, *European Journal of Operational Research*, *The Journal of the Operational Research Society*, *Journal of Productivity Analysis*, *Fuzzy Sets and Systems*, *International Journal of Information Technology and Decision Making*, *Top*, and others.

Kaoru Tone has been graduated from Department of Mathematics, the University of Tokyo, in 1953. After serving as an Associate Professor at Keio University, he joined Graduate School of Policy Science at Saitama University in 1977. Currently, he is a Professor Emeritus and Academic Fellow at National Graduate Institute for Policy Studies. He is an honorary member of the Operations Research Society of Japan and served as President of the Operations Research Society of Japan, 1996–1998. His research interests include theory and applications of optimizations, evaluation methods for managerial and technical efficiency of performances, and decision-making methods. He is conducting cooperative studies on the above subjects with Institute of Posts and Telecommunications Policy, National Land Agency and RIKEN, Japan. Also, he is on the *Editorial Boards of OMEGA*, *Socio-Economic Planning Sciences* (*International Journal of Public Sector Decision Making*), and *Encyclopedia of Operations Research and Management Science*, among others. His works on the above subjects have appeared in *European Journal of Operational Research*, *OMEGA*, *Socio-Economic Planning Sciences*, *Annals of Operations Research*, *Journal of Productivity Analysis*, *International Journal of Production Economics*, *Journal of the Operational Research Society*, *Journal of the Operational Research Society Japan*.

Konstantinos P. Triantis has been educated at Columbia University, where he received his BS and MS in Industrial and Management engineering, and M Phil and PhD in Industrial Engineering and Operations Research. He is a Professor in the Grado Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University, where he also serves as the Research Director of the System Performance Laboratory (SPL), the Director of the Systems Engineering Program, and the Academic Director of the of the Engineering Administration Program. He is currently serving as the Program Director of the Civil Infrastructure Systems Program at the National Science Foundation. His teaching and research interests cover the areas of performance measurement and evaluation, the design of

performance management systems, DEA, systems thinking and modeling, and fuzzy methods. His research has been published in a number of journals including *Management Science*, *European Journal of Operational Research*, *Journal of the Operational Research Society*, *Technometrics*, *IEEE Transactions on Engineering Management*, *Annals of Operations Research*, *Journal of Productivity Analysis*, *Transportation Research Parts A and E*, *Journal of Infrastructure Management*, *Transportation*, among others. He has received research funding from a number of organizations that include the National Science Foundation, Office of Naval Research, Department of Energy, American Red Cross, United States Postal Service, among others. He is a member of INFORMS, ASQ, System Dynamics Society, and IIE.

Paul W. Wilson is Professor of Economics at Clemson University in Clemson, South Carolina, where he specializes in econometrics and empirical microeconomics. His theoretical research interests include nonparametric estimation and inference, resampling methods, outlier detection, and computational methods. His empirical research includes studies of mergers, failures, efficiency, RTS, etc. in the banking and hospital industries, demand for health-care services, urban transportation, housing, and other issues. He has worked as a consultant for the U.S. Federal Reserve System, U.S. Department of Veterans Affairs, World Bank, and private industry. He is author of the widely used FEAR package for frontier efficiency analysis, and is currently editor of *Journal of Productivity Analysis*. Before moving to Clemson University in 2006, he was Professor of Economics at the University of Texas in Austin.

Zijiang Yang received her bachelor's degree in Computer Science from the Beijing Institute of Technology in China. Upon arrival in Canada she enrolled in the Graduate school at the University of Toronto where she received her MASc and PhD in Industrial Engineering. Her focus of research was on information technology and banking productivity, respectively. She is currently an Associate Professor in the School of Information Technology at York University, Toronto, Canada. Her research now includes work in operations research and artificial intelligence approaches for risk management in financial decision-making areas such as financial distress evaluation, credit risk assessment, e-commerce anomaly detection, and others. Her work includes the development of OR and artificial intelligence models for financial institutions.

Haiyan Zhu is a post-doctoral research fellow at the Centre for Management of Technology & Entrepreneurship, University of Toronto. Research areas of interest are applying numerical and analytical analysis knowledge and expertise to solve complex business problems pertaining to process optimization; performance measurement; productivity evaluation; and resource reallocation. Ms. Haiyan Zhu holds a Mechanical Engineering and MASc degrees from China and PhD and a Chemical Engineering from the University of Western Ontario.

Joe Zhu is Professor of Operations, School of Business at Worcester Polytechnic Institute, Worcester, MA. His research interests include issues of performance evaluation and benchmarking, and DEA. He has published over 80 articles in journals such as *Management Science*, *Operations Research, IIE Transactions*, *Naval Research Logistics*, *European Journal of Operational Research Annals of Operations Research*, *Journal of Operational Research Society*, *Computer and Operations Research*, *OMEGA*, *Journal of Productivity Analysis*, and others. He is the author of *Quantitative Models for Evaluating Business Operations: Data Envelopment Analysis with Spreadsheets, second edition* (Springer Science, 2009). He has also co-authored and co-edited books focusing on DEA. He developed the DEAFrontier software which is a DEA add-in for Microsoft Excel. He is an Area Editor of *OMEGA*, and on the editorial board of *Computers & Operations Research*. For information on his research, please visit www.deafrontier.net.

Contents

Contributors

Necmi K. Avkiran UQ Business School, The University of Queensland, Brisbane, QLD 4072, Australia
n.avkiran@business.uq.edu.au

Rajiv D. Banker Fox School of Business and Management, Temple University, Philadelphia, PA 19122, USA
banker@temple.edu

Jon A. Chilingerian Heller School for Social Policy and Management, Brandeis University Waltham, Waltham, MA 02254-9110, USA
Chilingerian@Brandeis.edu

Wade D. Cook Schulich School of Business, York University, Toronto, Canada, M3J 1P3
wcook@schulich.yorku.ca

William W. Cooper Red McCombs School of Business, University of Texas at Austin, Austin, TX 78712, USA
cooperw@mail.utexas.edu

Honghui Deng School of Business, University of Nevada, Las Vegas, NV 89154, USA
honghui.deng@unlv.edu

Rolf Färe Department of Economics, Oregon State University, Corvallis, OR 97331, USA

Shawna Grosskopf Department of Economics, Oregon State University, Corvallis, OR 97331, USA

Zhimin Huang School of Business, Adelphi University, Garden City, NY 11530, USA
huang@adelphi.edu

Shanling Li Faculty of Management, McGill University,
1001 Sherbrooke Street West, Montreal, QC H3A 1G5, Canada
shanling.li@mcgill.ca

Susan X. Li School of Business, Adelphi University, Garden City,
NY 11530, USA
li@adelphi.edu

Liang Liang School of Management, University of Science and Technology
of China, He Fei, An Hui Province, People's Republic of China
lliang@ustc.edu.cn

Dimitris Margaritis Department of Accounting & Finance, University
of Auckland Business School, Auckland, New Zealand

Ram Natarajan School of Management, The University of Texas at Dallas,
Richardson, TX 75083-0688, USA
nataraj@utdallas.edu

Joseph C. Paradi Centre for Management of Technology and Entrepreneurship,
University of Toronto, Toronto, ON, Canada
paradi@mie.utoronto.ca

José L. Ruiz Centro de Investigación Operativa, Universidad Miguel Hernández,
Avd. de la Universidad, s/n 03202, Elche (Alicante), Spain
jlruiz@umh.es

Lawrence M. Seiford Department of Industrial and Operations Engineering,
University of Michigan at Ann Arbor, Ann Arbor, MI 48102, USA
seiford@umich.edu

H. David Sherman College of Business Administration, Northeastern
University, Boston, MA 02115, USA
H.Sherman@NEU.edu

Léopold Simar Institut de Statistique and CORE, Université Catholique
de Louvain, Voie du Roman Pays 20, Louvain-la-Neuve, Belgium
simar@stat.ucl.ac.be

Inmaculada Sirvent Centro de Investigación Operativa, Universidad Miguel
Hernndez, Avd. de la Universidad, s/n 03202, Elche (Alicante), Spain

Kaoru Tone National Graduate Institute for Policy Studies, 7-22-1 Roppongi,
Minato-ku, Tokyo 106-8677, Japan
tone@grips.ac.jp

Konstantinos P. Triantis System Performance Laboratory, Grado Department
of Industrial and Systems Engineering, Northern Virginia Center, Virginia Tech,
7053 Haycock Road, Falls Church, VA 22043-2311, USA
triantis@vt.edu

Paul W. Wilson Department of Economics, Clemson University, Clemson, SC, USA
pww@clemson.edu

Zijiang Yang School of Information Technology, York University, 4700 Keele Street, Toronto, ON, Canada

Haiyan Zhu Centre for Management of Technology and Entrepreneurship, University of Toronto, Toronto, ON, Canada

Joe Zhu School of Business, Worcester Polytechnic Institute, Worcester, MA 01609, USA
jzhu@wpi.edu

Chapter 1
Data Envelopment Analysis: History, Models, and Interpretations*

William W. Cooper, Lawrence M. Seiford, and Joe Zhu

Abstract In about 30 years, Data Envelopment Analysis (DEA) has grown into a powerful quantitative, analytical tool for measuring and evaluating the performance. DEA has been successfully applied to a host of many different types of entities engaged in a wide variety of activities in many contexts worldwide. This chapter discusses the basic DEA models and some of their extensions.

Keywords Data envelopment analysis • Efficiency • Performance

1.1 Introduction

Data Envelopment Analysis (DEA) is a "data-oriented" approach for evaluating the performance of a set of peer entities called Decision-Making Units (DMUs), which convert multiple inputs into 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

*Part of the material in this chapter is adapted from the Journal of Econometrics, Vol. 46, Seiford, L.M. and Thrall, R.M., Recent developments in DEA: The mathematical programming approach to frontier analysis, 7–38, 1990, with permission from Elsevier Science.

J. Zhu (✉)
School of Business, Worcester Polytechnic Institute, Worcester, MA 01609, USA
e-mail: jzhu@wpi.edu

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_1,

for use in cases that 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.

As pointed out in Cooper et al. (2007), DEA has also been used to supply new insights into activities (and entities) that have previously been evaluated by other methods. For instance, studies of benchmarking practices with DEA have identified numerous sources of inefficiency in some of the most profitable firms – firms that had served as benchmarks by reference to this (profitability) criterion – but DEA has provided a vehicle for identifying better benchmarks in many applied studies. Because of these possibilities, DEA studies of the efficiency of different legal organization forms such as "stock" vs. "mutual" insurance companies have shown that previous studies have fallen short in their attempts to evaluate the potentials of these different forms of organizations. Similarly, a use of DEA has suggested reconsideration of previous studies of the efficiency with which pre-and postmerger activities have been conducted in banks that were studied by DEA.

Since DEA was first introduced in 1978 in its present form, 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. This has been accompanied by other developments. For instance, Zhu (2003a, 2009) 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. See, for instance, the use of DEA to guide removal of the Diet and other government agencies from Tokyo as described in Takamura and Tone (2003).

In their originating article, Charnes et al. (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 regressions, 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 would 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 explicitly formulated assumptions and variations that are required with various types of models such as linear and nonlinear regression models.

Relative efficiency in DEA accords with the following definition, which has the advantage of avoiding the need for assigning a priori measures of relative importance to any input or output,

Definition 1.1 (Efficiency – Extended Pareto-Koopmans Definition). Full (100%) efficiency is attained by any DMU if and only if none of its inputs or outputs can be improved without worsening some of its other inputs or outputs.

In most management or social science applications, the theoretically possible levels of efficiency will not be known. The preceding definition is, therefore, replaced by emphasizing its uses with only the information that is empirically available as in the following definition:

Definition 1.2 (Relative Efficiency). A DMU is to be rated as fully (100%) efficient on the basis of available evidence if and only if the performances of other DMUs does not show that some of its inputs or outputs can be improved without worsening some of its other inputs or outputs.

Notice that this definition avoids the need for recourse to prices or other assumptions of weights, which are selected a priori and are supposed to reflect the relative importance of the different inputs or outputs. It also avoids the need for explicitly specifying the formal relations that are supposed to exist between inputs and outputs. This basic kind of efficiency, referred to as "technical efficiency" in economics can, however, be extended to other kinds of efficiency when data such as prices, unit costs, etc., are available for use with DEA.

In this chapter, we discuss the mathematical programming approach of DEA that implements the above efficiency definition. Section 1.2 of this chapter provides a historical perspective on the origins of DEA. Section 1.3 provides a description of the original "CCR ratio model" of Charnes, Cooper, and Rhodes, which relates the above efficiency definition to other definitions of efficiency such as the ones used in engineering and science, as well as in business and economics. Section 1.4 describes some methodological extensions that have been proposed. Section 1.5 expands the development to concepts such as "allocative" (or price) efficiency, which can add additional power to DEA when unit prices and costs are available. This is done in Sect. 1.5 and extended to profit efficiency in Sect. 1.6 after which some recent DEA developments are discussed in Sect. 1.7 and a conclusion in Sect. 1.8 is supplied.

1.2 Background and History

In an article that represents the inception of DEA, Farrell (1957) was motivated by the need for developing better methods and models for evaluating productivity. He argued that while attempts to solve the problem usually produced careful measurements, they were also very restrictive because they failed to combine the measurements of multiple inputs into any satisfactory overall measure of efficiency. Responding to these inadequacies of separate indices of labor productivity, capital productivity, etc., Farrell proposed an activity analysis approach that could more

adequately deal with the problem. His measures were intended to be applicable to any productive organization; in his words, "... from a workshop to a whole economy." In the process, he extended the concept of "productivity" to the more general concept of "efficiency."

Our focus in this chapter is on basic DEA models for measuring the efficiency of a DMU *relative* to similar DMUs to estimate a "best practice" frontier. The initial DEA model, as originally presented in Charnes, Cooper, and Rhodes (CCR), was built on the earlier work of Farrell.

This work by Charnes, Cooper, and Rhodes originated in the early 1970s in response to the thesis efforts of Edwardo Rhodes at Carnegie Mellon University's School of Urban & Public Affairs – now the H.J. Heinz III School of Public Policy and Management. Under the supervision of W.W. Cooper, this thesis was to be directed to evaluating educational programs for disadvantaged students (mainly black or Hispanic) in a series of large scale studies undertaken in US public schools with support from the Federal government. Attention was finally centered on Program Follow Through - a huge attempt by the US Office (now Department) of Education to apply principles from the statistical design of experiments to a set of matched schools in a nation-wide study. Rhodes secured access to the data being processed for that study by Abt Associates, a Boston-based consulting film, under contract with the US Office of Education. The database was sufficiently large so that issues of degrees of freedom, etc., were not a serious problem despite the numerous input and output variables used in the study. Nevertheless, unsatisfactory and even absurd results were secured from all of the statistical-econometric approaches that Rhodes attempted to use.

While trying to respond to this situation, Rhodes called Cooper's attention to Farrell's seminal article. In this chapter, Farrell used "activity analysis concepts" to correct what he believed were deficiencies in commonly used index number approaches to productivity (and like) measurements.

Cooper had previously worked with A. Charnes to give computationally implementable form to Tjalling Koopmans' "activity analysis concepts." So, taking Farrell's statements at face value, Cooper and Rhodes formalized what was involved in the definitions that were given in Sect. 1.1 of this chapter. These definitions then provided the guides that were used for their subsequent research.

The name of Pareto is assigned to the first of these two definitions for the following reasons. In his *Manual of Political Economy* (1906), the Swiss-Italian economist Vilfredo Pareto established the basis of modern "welfare economics," i.e., the part of economics concerned with evaluating public policies, by noting that a social policy could be justified if it made some persons better off without making others worse off. In this way, the need for making comparisons between the value of the gains to some and the losses to others could be avoided. This avoids the necessity of ascertaining the "utility functions" of the affected individuals and/or to "weight" the relative importance of each individual's gains and losses.

This property, known as the "Pareto criterion" as used in welfare economics, was carried over, or adapted, in a book edited by Koopmans (1951). In this context, it was "final goods" that were accorded this property, in that they were all

constrained so that no final good was allowed to be improved if this improvement resulted in worsening one or more other final goods. These final goods (=outputs) were to be satisfied in stipulated amounts, while inputs were to be optimally determined in response to the prices and amounts exogenously fixed for each output (=final good). Special attention was then directed by Koopmans to "efficiency prices," which are the prices associated with efficient allocation of resources (=inputs) to satisfy the preassigned demands for final goods. For a succinct summary of the mechanisms involved in this "activity analysis" approach, see p. 299 in Charnes and Cooper (1961).

Pareto and Koopmans were concerned with analyses of entire economies. In such a context, it is reasonable to allow input prices and quantities to be determined by reference to their ability to satisfy final demands. Farrell, however, extended the Pareto-Koopmans property to inputs as well as outputs and explicitly eschewed any use of prices and/or related "exchange mechanisms." Even more importantly, he used the performances of other DMUs to evaluate the behavior of each DMU relative to the outputs and the inputs each of the other DMUs used. This made it possible to empirically determine their relative efficiencies.

The resulting measure which is referred to as the "Farrell measure of efficiency," was regarded by Farrell as restricted to meaning "technical efficiency" or the amount of "waste" that can be eliminated without worsening any input or output. This was then distinguished by Farrell from "allocative" and "scale" efficiencies as adapted from the literature of economics. These additional efficiencies are discussed later in this chapter where the extensions needed to deal with problems that were encountered in DEA attempts to use these concepts in actual applications are also discussed. Here, we want to note that Farrell's approach to efficiency evaluations, as embodied in the "Farrell measure," carries with it an assumption of equal access to inputs by all DMUs. This does not mean that all DMUs use the same input amounts, however, and, indeed, part of their efficiency evaluations will depend on the input amounts used by each DMU as well as the outputs that they produce.

This "equal access assumption" is a mild one, at least as far as data availability is concerned. It is less demanding than the data and other requirements needed to deal with aspects of performance such as "allocative" or "scope" and "scale efficiencies." Furthermore, as discussed below, this assumption can now be relaxed. For instance, one can introduce "nondiscretionary variables and constraints" to deal with conditions beyond the control of a DMUs management – in the form of "exogenously" fixed resources (such as weather), which may differ for each DMU. One can also introduce "categorical variables" to insure that evaluations are effected by reference to DMUs that have similar characteristics, and still other extensions and relaxations are possible, as covered in the discussions that follow.

The definition of efficiency that we have referred to as "Extended Pareto-Koopmans Efficiency" and "Relative Efficiency" were formalized by Charnes, Cooper, and Rhodes rather than Farrell. However, these definitions conform both to Farrell's models and the way Farrell used them. In any case, these were the definitions that Charnes, Cooper, and Rhodes used to guide the developments that we next describe.

The Program Follow Through data with which Rhodes was concerned in his thesis recorded "outputs" such as "increased self esteem in a disadvantaged child" and "inputs" such as "time spent by a mother in reading with her child," as measured by psychological tests and prescribed record keeping and reporting practices. Farrell's elimination of the need for information on prices proved attractive for dealing with outputs and inputs such as these – as reported for each of the schools included in the Program Follow Through experiment.

Farrell's empirical work had been confined to single-output cases and his sketch of extensions to multiple outputs did not supply what was required for applications to large datasets such as those involved in Program Follow Through. To obtain what was needed in computationally implementable form, Charnes, Cooper, and Rhodes developed the dual pair of linear programming problems that are modeled in the next section, Sect. 1.3. It was then noticed that Farrell's measure failed to account for the nonzero slacks, which is where the changes in proportions connected with mix inefficiencies are located (in both outputs and inputs). The possible presence of nonzero slack as a source of these mix inefficiencies also requires attention even when restricted to "technical efficiency."

We now emphasize the problems involved in dealing with these slacks because a considerable part of the DEA (and related) literatures continues to be deficient in its treatment of nonzero slack. A significant part of the problem to be dealt with, as we noted above, involves the possible presence of alternate optima in which the same value of the Farrell measure could be associated with zero slack in some optima but not in others. Farrell introduced "points at infinity" in what appears to have been an attempt to deal with this problem but was unable to give operationally implementable form to this concept. Help in dealing with this problem was also not available from the earlier work of Sidney Afriat (1972), Ronald Shephard (1970), or Gerhard Debreu (1951). To address this problem, Charnes, Cooper, and Rhodes introduced mathematical concepts that are built around the "non-Archimedean" elements associated with $\varepsilon > 0$, which handles the problem by insuring that slacks are always maximized without altering the value of the Farrell measure.

The dual problems devised by Cooper and Rhodes readily extended the above ideas to multiple outputs and multiple inputs in ways that could locate inefficiencies in each input and each output for every one of the DMUs. Something more was nevertheless desired in the way of summary measures. At this point, Cooper invited A. Charnes to join him and Rhodes in what promised to be a very productive line of research. Utilizing the earlier work of Charnes and Cooper (1962), which had established the field of "fractional programming," Charnes was able to put the dual linear programming problems devised by Cooper and Rhodes into the equivalent ratio form represented in (1.1) below and this provided a basis for unifying what had been done in DEA with long-standing approaches to efficiency evaluation and analysis used in other fields, such as engineering and economics.

Since the initial study by Charnes, Cooper, and Rhodes some 4,000 articles have appeared in the literature. See Emrouznejad et al. (2008). Such rapid growth and widespread (and almost immediate) acceptance of the methodology of DEA is

testimony to its strengths and applicability. Researchers in a number of fields have quickly recognized that DEA is an excellent methodology for modeling operational processes, and its empirical orientation and minimization of a priori assumptions have resulted in its use in a number of studies involving efficient frontier estimation in the nonprofit sector, in the regulated sector, and in the private sector.

At present, DEA actually encompasses a variety of alternate (but related) approaches to evaluating performance. Extensions to the original CCR work have resulted in a deeper analysis of both the "multiplier side" from the dual model and the "envelopment side" from the primal model of the mathematical duality structure. Properties such as isotonicity, nonconcavity, economies of scale, piecewise linearity, Cobb–Douglas loglinear forms, discretionary and nondiscretionary inputs, categorical variables, and ordinal relationships can also be treated through DEA. Actually the concept of a frontier is more general than the concept of a "production function" which has been regarded as fundamental in economics in that the frontier concept admits the possibility of multiple production functions, one for each DMU, with the frontier boundaries consisting of "supports" that are "tangential" to the more efficient members of the set of such frontiers.

1.3 CCR Model

To allow for applications to a wide variety of activities, we use the term Decision-Making Unit (DMU) to refer to any entity that is to be evaluated in terms of its abilities to convert inputs into outputs. These evaluations can involve governmental agencies and not-for-profit organizations as well as business firms. The evaluation can also be directed to educational institutions and hospitals as well as police forces (or subdivision thereof) or army units for which comparative evaluations of their performance are to be made. See Bessent et al. (1983).

We assume that there are n DMUs to be evaluated. Each DMU consumes varying amounts of m different inputs to produce s different outputs. Specifically, DMU_j consumes amount x_{ij} of input i and produces amount y_{rj} of output r. We assume that $x_{ij} \geq 0$ and $y_{rj} \geq 0$ and further assume that each DMU has at least one positive input and one positive output value.

We now turn to the "ratio-form" of DEA. In this form, as introduced by Charnes, Cooper, and Rhodes, the ratio of outputs to inputs is used to measure the relative efficiency of the $\text{DMU}_j = \text{DMU}_o$ to be evaluated relative to the ratios of all of the $j = 1, 2, \ldots, n$ DMU_j. We can interpret the CCR construction as the reduction of the multiple-output/multiple-input situation (for each DMU) to that of a single "virtual" output and "virtual" input. For a particular DMU the ratio of this single virtual output to single virtual input provides a measure of efficiency that is a function of the multipliers. In mathematical programming parlance, this ratio, which is to be maximized, forms the objective function for the particular DMU being evaluated, so that symbolically

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

where it should be noted that the variables are the u_rs and the v_is and the y_{ro}s and x_{io}s are the observed output and input values, respectively, of 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 $\text{DMU}_j = \text{DMU}_o$, must be less than or equal to unity. The mathematical programming problem may thus be stated as

$$\max h_o(u, v) = \frac{\sum_r u_r y_{ro}}{\sum_i v_i x_{io}} \tag{1.2}$$

subject to

$$\frac{\sum_r u_r y_{rj}}{\sum_i v_i x_{ij}} \leq 1 \text{ for } j = 1, \ldots, n,$$

$$u_r, v_i \geq 0 \text{ for all } i \text{ and } r.$$

This ratio form generalizes the engineering science definition of efficiency from a single output to a single input and does so without requiring the use of a priori chosen weights. See Bulla et al. 2000, for an application to the evaluation of jet aircraft engines.

Remark. A fully rigorous development would replace $u_r, v_i \geq 0$ with $\frac{u_r}{\sum_{i=1}^{m} v_i x_{io}}, \frac{u_r}{\sum_{i=1}^{m} v_i x_{io}} \geq \varepsilon > 0$ where ε is a non-Archimedean element smaller than any positive real number. See Arnold et al. (1998). This condition guarantees that solutions will be positive in these variables. It also leads to the $\varepsilon > 0$ in (1.6) which, in turn, leads to the second stage optimization of the slacks as in (1.10). It should be noted that (1.1) and (1.2) generalize the engineering-science definition of efficiency which deals with a single output-to-a-single-input ratio and requires an a priori assumption of weighting values to deal with multiple outputs and inputs. See Bulla et al. (2000) for an application of DEA to evaluate jet engines to be used in aircraft.

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

$$\begin{aligned}
&\max\ z = \sum_{r=1}^{s} \mu_r y_{ro} \\
&\text{subject to} \\
&\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0 \\
&\sum_{i=1}^{m} v_i x_{io} = 1 \\
&\mu_r, v_i \ge 0
\end{aligned} \tag{1.3}$$

for which the LP dual problem is

$$\begin{aligned}
&\theta^* = \min \theta \\
&\text{subject to} \\
&\sum_{j=1}^{n} x_{ij}\lambda_j \le \theta x_{io} \quad i = 1, 2, \ldots, m; \\
&\sum_{j=1}^{n} y_{rj}\lambda_j \ge y_{ro} \quad r = 1, 2, \ldots, s; \\
&\lambda_j \ge 0 \qquad\qquad j = 1, 2, \ldots, n.
\end{aligned} \tag{1.4}$$

This last model, (1.4), is sometimes referred to as the "Farrell model" because it is the one used in Farrell. In the economics portion of the DEA literature, it is said to conform to the assumption of "strong disposal," but the efficiency evaluation it makes ignores the presence of nonzero slacks. In the operations research portion of the DEA literature, this is referred to as "weak efficiency."

Possibly because he used the literature of "activity analysis" for reference – see Koopmans – Farrell also failed to exploit the very powerful dual theorem of linear programming, which we have used to relate the preceding problems to each other. This use of activity analysis also caused computational difficulties for Farrell because he did not take advantage of the fact that activity analysis models can be converted to linear programming equivalent that provide immediate access to the simplex and other methods for efficiently solving such problems. See, e.g., Charnes and Cooper (1961 Chap. 9). We, therefore, now begin to bring these features of linear programming into play.

By virtue of the dual theorem of linear programming, we have $z^* = \theta^*$. Hence, either problem may be used. One can solve say (1.4), to obtain an efficiency score. Because we can set $\theta = 1$ and $\lambda_k^* = 1$ with $\lambda_k^* = \lambda_o^*$ and all other $\lambda_j^* = 0$, a solution of (1.4) always exists. Moreover, this solution implies $\theta^* \le 1$. The optimal solution, θ^*, yields an efficiency score for a particular DMU. The process is repeated for each DMU_j, i.e., solve (1.4), with $(X_o, Y_o) = (X_k, Y_k)$, where (X_k, Y_k) represent vectors

with components x_{ik}, y_{rk} and, similarly (X_o, Y_o) has components x_{ok}, y_{ok}. DMUs for which $\theta^* < 1$ are inefficient, while DMUs for which $\theta^* = 1$ are boundary points.

Some boundary points may be "weakly efficient" because we have nonzero slacks. This may appear to be worrisome because alternate optima may have nonzero slacks in some solutions, but not in others. However, we can avoid being worried even in such cases by invoking the following linear program in which the slacks are taken to their maximal values.

$$\begin{aligned}
&\max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \\
&\text{subject to} \\
&\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta^* x_{io} \quad i = 1, 2, \ldots, m; \\
&\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = y_{ro} \quad r = 1, 2, \ldots, s; \\
&\lambda_j, s_i^-, s_r^+ \geq 0 \forall i, j, r
\end{aligned} \tag{1.5}$$

where we note the choices of s_i^- and s_r^+ do not affect the optimal θ^*, which is determined from model (1.4).

These developments now lead to the following definition based upon the "relative efficiency" definition 1.2, which was given in Sect. 1.1 above.

Definition 1.3 (DEA Efficiency). The performance of DMU_o is fully (100%) efficient if and only if both (1) $\theta^* = 1$ and (2) all slacks $s_i^{-*} = s_r^{+*} = 0$.

Definition 1.4 (Weakly DEA Efficient). The performance of DMU_o is weakly efficient if and only if both (1) $\theta^* = 1$ and (2) $s_i^{-*} \neq 0$ and/or $s_r^{+*} \neq 0$ for some i or r in some alternate optima.

It is to be noted that the preceding development amounts to solving the following problem in two steps:

$$\begin{aligned}
&\min \theta - \varepsilon \left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right) \\
&\text{subject to} \\
&\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta x_{io} \quad i = 1, 2, \ldots, m; \\
&\sum_{j=1}^{n} y_{rj}\lambda_j - s_i^+ = y_{ro} \quad r = 1, 2, \ldots, s; \\
&\lambda_j, s_i^-, s_r^+ \geq 0 \, \forall i, j, r
\end{aligned} \tag{1.6}$$

where the s_i^- and s_r^+ are slack variables used to convert the inequalities in (1.4) to equivalent equations. Here, $\varepsilon > 0$ is a so-called non-Archimedean element defined

to be smaller than *any* positive real number. This is equivalent to solving (1.4) in two stages by first minimizing θ, then fixing $\theta = \theta^*$ as in (1.2), where the slacks are to be maximized without altering the previously determined value of $\theta = \theta^*$. Formally, this is equivalent to granting "preemptive priority" to the determination of θ^* in (1.3). In this manner, the fact that the non-Archimedean element ε is defined to be smaller than any positive real number is accommodated without having to specify the value of ε.

Alternately, one could have started with the output side and considered instead the ratio of virtual input to output. This would reorient the objective from max to min, as in (1.2), to obtain

$$\begin{aligned}
&\text{Min} \frac{\sum_i v_i x_{io}}{\sum_r u_r y_{ro}} \\
&\text{Subject to} \\
&\frac{\sum_i v_i x_{ij}}{\sum_r u_r y_{rj}} \geq 1 \text{ for } j = 1, \ldots, n, \\
&u_r, v_i \geq \varepsilon > 0 \text{ for all } i \text{ and } r
\end{aligned} \tag{1.7}$$

where $\varepsilon > 0$ is the previously defined non-Archimedean element.

Again, the Charnes and Cooper (1962) transformation for linear fractional programming yields model (1.8) (multiplier model) below, with associated dual problem, (1.9) (envelopment model), as in the following pair,

$$\begin{aligned}
&\min\ q = \sum_{i=1}^{m} v_i x_{io} \\
&\text{subject to} \\
&\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} \geq 0 \\
&\sum_{r=1}^{s} \mu_r y_{ro} = 1 \\
&\mu_r, v_i \geq \varepsilon, \quad \forall r,\ i
\end{aligned} \tag{1.8}$$

$$\begin{aligned}
&\max\ \varphi + \varepsilon \left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right) \\
&\text{subject to} \\
&\sum_{j=1}^{n} x_{ij} \lambda_j + s_i^- = x_{io} \qquad i = 1, 2, \ldots, m; \\
&\sum_{j=1}^{n} y_{rj} \lambda_j - s_r^+ = \varphi y_{ro} \quad r = 1, 2, \ldots, s; \\
&\lambda_j \geq 0 \qquad\qquad\qquad\quad j = 1, 2, \ldots, n.
\end{aligned} \tag{1.9}$$

See pp. 75–76 in Cooper et al. (2007) for a formal development of this transformation and modification of the expression for $\varepsilon > 0$. See also the remark following (1.2).

Here, we use a model with an output-oriented objective as contrasted with the input orientation in (1.6). However, as before, model (1.9) 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,

$$\begin{aligned}
&\max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \\
&\text{subject to} \\
&\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = x_{io} \qquad i = 1, 2, \ldots, m; \\
&\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = \varphi^* y_{ro} \qquad r = 1, 2, \ldots, s; \\
&\lambda_j \geq 0 \qquad\qquad\qquad\qquad j = 1, 2, \ldots, n. \qquad (1.10)
\end{aligned}$$

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

Definition 1.5. DMU_o is efficient if and only if $\varphi^* = 1$ and $s_i^{-*} = s_r^{+*} = 0$ for all i and r. DMU_o is weakly efficient if $\varphi^* = 1$ and $s_i^{-*} \neq 0$ and (or) $s_r^{+*} \neq 0$ for some i and r in some alternate optima.

Table 1.1 presents the CCR model in input- and output-oriented versions, each in the form of a pair of dual linear programs.

These are known as CCR (Charnes et al. 1978) models. If the constraint $\sum_{j=1}^{n} \lambda_j = 1$ is adjoined, they are known as BCC (Banker et al. 1984) models. This added constraint introduces an additional variable, μ_o, into the (dual) multiplier problems. As seen in the next chapter, this extra variable makes it possible to effect returns-to-scale evaluations (increasing, constant, and decreasing). Thus, the BCC model is also referred to as the VRS (Variable Returns to Scale) model and distinguished from the CCR model which is referred to as the CRS (Constant Returns to Scale) model.

We now proceed to compare and contrast the input and output orientations of the CCR model. To illustrate the discussion to be followed, we employ the example presented in Fig. 1.1 consisting of five DMUs, labeled P1, . . ., P5, each consuming a single input to produce a single output.

The numbers in parentheses in Fig. 1.1 are interpreted as coordinate values which correspond to the input and output of a DMU_j represented as Pj, $j = 1, \ldots, 5$. In each case the value on the left in the parentheses alongside the point is the point's input and the value on the right is the output for the Pj alongside which these values are listed.

Table 1.1 CCR DEA model

Input-oriented	
Envelopment model	Multiplier model
$\min \theta - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)$	$\max z = \sum_{r=1}^{s} \mu_r y_{ro}$
subject to	subject to
$\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta x_{io} \quad i = 1, 2, \ldots, m;$	$\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0$
$\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = y_{ro} \quad r = 1, 2, \ldots, s;$	$\sum_{i=1}^{m} v_i x_{io} = 1$
$\lambda_j \geq 0 \quad j = 1, 2, \ldots, n$	$\mu_r, v_i \geq \varepsilon > 0$
Output-oriented	
Envelopment model	Multiplier model
$\max \varphi + \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)$	$\min q = \sum_{i=1}^{m} v_i x_{io}$
subject to	subject to
$\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = x_{io} \quad i = 1, 2, \ldots, m;$	$\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} \geq 0$
$\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = \varphi y_{ro} \quad r = 1, 2, \ldots, s;$	$\sum_{r=1}^{s} \mu_r y_{ro} = 1$
$\lambda_j \geq 0 \quad j = 1, 2, \ldots, n.$	$\mu_r, v_i \geq \varepsilon > 0$

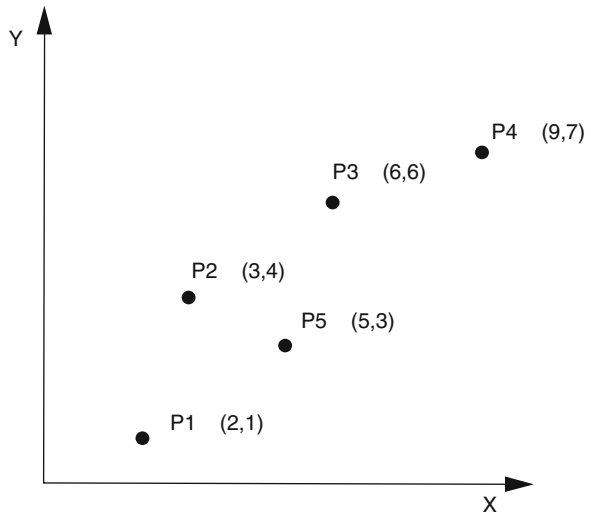


Fig. 1.1 Example DMUs

Table 1.2 Optimal solution values for the CCR model

	DMU	z^*	μ^*	v^*	θ^*	λ^*
Input-oriented	1	3/8	3/8	1/2	3/8	$\lambda_2 = 1/4$
	2	1	¼	1/3	1	$\lambda_2 = 1$
	3	¾	1/8	1/6	¾	$\lambda_2 = 3/2$
	4	7/12	1/12	1/9	7/12	$\lambda_2 = 7/4$
	5	9/20	3/20	1/5	9/20	$\lambda_2 = 3/4$
Output-oriented	1	8/3	1	4/3	8/3	$\lambda_2 = 2/3$
	2	1	¼	1/3	1	$\lambda_2 = 1$
	3	4/3	1/6	2/9	4/3	$\lambda_2 = 2$
	4	12/7	1/7	4/21	12/7	$\lambda_2 = 3$
	5	20/9	1/3	4/9	20/9	$\lambda_2 = 5/3$

To assist the reader in verifying the model interpretations which follow, Table 1.2 contains optimal solution values for the five example DMUs for both of the dual LP problems of the CCR model. For example, to evaluate the efficiency of P5 (DMU_5 in Table 1.2), we can solve the following input-oriented envelopment CCR model:

$$
\begin{aligned}
&\min\theta \\
&\text{subject to} \\
&2\lambda_1 + 3\lambda_2 + 6\lambda_3 + 9\lambda_4 + 5\lambda_5 \leq 5\theta \ \text{(input)} \\
&1\lambda_1 + 4\lambda_2 + 6\lambda_3 + 7\lambda_4 + 3\lambda_5 \geq 3 \ \text{(output)} \\
&\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 \geq 0
\end{aligned}
$$

which yields the values of $\theta^* = 9/20$, $\lambda_2^* = 3/3$, and $\lambda_j^* = 0$ ($j \neq 2$) (see the last two columns in the last row of the upper portion of Table 1.2).

Alternatively, we can solve the input-oriented multiplier CCR model,

$$
\begin{aligned}
&\max\ z = 3\mu \\
&\text{subject to} \\
&1\mu - 2v \leq 0 \ \text{(P1)} \\
&4\mu - 3v \leq 0 \ \text{(P2)} \\
&6\mu - 6v \leq 0 \ \text{(P3)} \\
&7\mu - 9v \leq 0 \ \text{(P4)} \\
&3\mu - 5v \leq 0 \ \text{(P5)} \\
&5v = 1 \\
&\mu, v \geq 0
\end{aligned}
$$

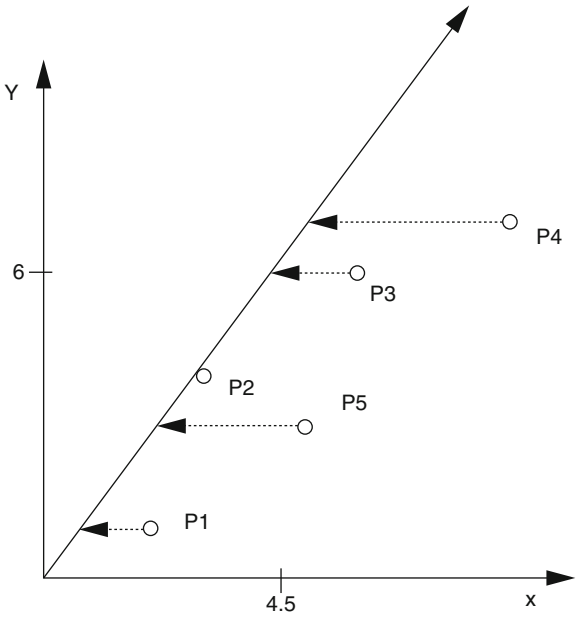


Fig. 1.2 Projection to frontier for the input-oriented CCR model

which yields $z^* = 9/20$, $\mu^* = 3/20$, and $v^* = 1/5$. Hence, we have $\theta^* = z^*$. Moreover, with $\mu^* = 3/20$ and $v^* = 1/5$, this is also the value of $h_o(u^*, v^*) = 3(\frac{3}{20})/5(\frac{1}{5}) = \frac{9}{20}$ for the corresponding ratio model obtained from (1.2).

A DMU is inefficient if the efficiency score given by the optimal value for the LP problem is less than one ($\theta^* < 1$ or $z^* < 1$). If the optimal value is equal to one *and* if there exist positive optimal multipliers ($\mu_r > 0$, $v_i > 0$), then the DMU is efficient. Thus, all efficient points lie on the frontier. However, a DMU can be a boundary point ($\theta^* = 1$) and be inefficient. Note that the complementary slackness condition of linear programming yields a condition for efficiency which is equivalent to the above; the constraints involving X_0 and Y_0, must hold with equality, i.e., $X_0 = X\lambda^*$ and $Y_0 = Y\lambda^*$ for all optimal λ^*, where X_0 and Y_0 are vectors and X and Y are matrices.

An inefficient DMU can be made more efficient by projection onto the frontier. In an input orientation one improves efficiency through proportional reduction of inputs, whereas an output orientation requires proportional augmentation of outputs. However, it is necessary to distinguish between a boundary point and an efficient boundary point. Moreover, the efficiency of a boundary point can be dependent upon the model orientation.

The efficient frontier and DEA projections are provided in Figs. 1.2 and 1.3 for the input-oriented and output-oriented CCR models, respectively. In both cases, the efficient frontier obtained from the CCR model is the ray $\{\alpha(x_2, y_2) | \alpha \geq 0\}$, where x_2 and y_2 are the coordinates of P2.

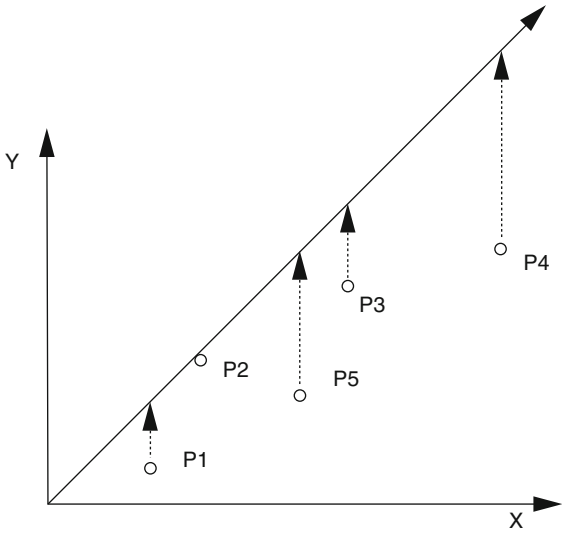


Fig. 1.3 Projection to frontier for the output-oriented CCR model

As can be seen from the points designated by the arrow head, an inefficient DMU may be projected to different points on the frontier under the two orientations. However, the following theorem provides a correspondence between solutions for the two models.

Theorem 1.1. Let (θ^*, λ^*) be an optimal solution for the input-oriented model in (1.9). Then, $((1/\theta*, \lambda * /\theta) = (\varphi*, \hat{\lambda}*)$ is optimal for the corresponding output-oriented model. Similarly if $(\varphi^*, \hat{\lambda}^*)$ is optimal for the output-oriented model then $(1/\varphi*, \hat{\lambda} * /\varphi*) = (\theta*, \lambda*)$ is optimal for the input-oriented model. The correspondence need not be 1-1, however, because of the possible presence of alternate optima.

For an input orientation, the projection $(X_0, Y_0) \rightarrow (\theta^* X_0, Y_0)$ always yields a boundary point. But technical efficiency is achieved only if also all slacks are zero in all alternate optima so that $\theta^* X_0 = X\lambda^*$ and $Y_0 = Y\lambda^*$ for all optimal λ^*. Similarly, the output-oriented projection $(X_0, Y_0) \rightarrow (X_0\varphi^*, Y_0)$ yields a boundary point which is efficient (technically) only if $\varphi^* Y_0 = Y\lambda^*$ and $X_0 = X\lambda^*$ for all optimal λ^*. That is, the constraints are satisfied as equalities in all alternate optima for (1.4). To achieve technical efficiency the appropriate set of constraints in the CCR model must hold with equality.

To illustrate this, we consider a simple numerical example used in Zhu (2009) as shown in Table 1.3 where we have five DMUs representing five 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.

We now turn to the BCC model for which Fig. 1.4 presents the five DMUs and the piecewise linear DEA frontier. DMUs 1, 2, 3, and 4 are on the frontier.

Table 1.3 Supply chain operations within a week

	Inputs		Output
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

Source: Zhu (2009)

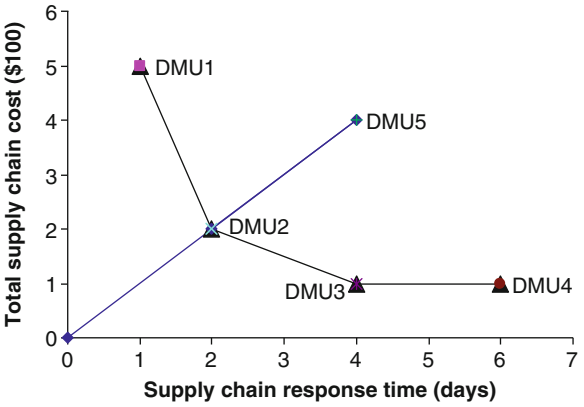


Fig. 1.4 Five supply chain operations. Source: Zhu (2009)

If we adjoin the constraint $\sum_{j=1}^{n} \lambda_j = 1$ to model (1.4) for DMU_5, we get from the data of Table 1.3,

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

This model has the unique optimal solution of $\theta^* = 0.5$, $\lambda_2^* = 1$, and $\lambda_j^* = 0$ ($j \neq 2$), indicating that DMU_5 needs to reduce its cost and response time to the amounts used by DMU_2 if it is to be efficient This example indicates that technical efficiency for DMU_5 is achieved at DMU_2 on the boundary.

Now, if we similarly use model (1.4) with $\sum_{j=1}^{n} \lambda_j = 1$ for DMU_4, we obtain$\theta^* = 1$, $\lambda_4^* = 1$, and $\lambda_j^* = 0$ ($j \neq 4$), indicating that DMU_4 is on the frontier and is a boundary point. However, Fig. 1.4 indicates that DMU_4 can still reduce its response time by 2 days to achieve coincidence with DMU_3 on the efficiency frontier.

This input reduction is the input slack, and the constraint with which it is associated is satisfied as a strict inequality in this solution. Hence, DMU_4 is weakly efficient and DMU_3 serves as a "benchmark" to improve the performance of DMU_4.

The nonzero slack can be found by using model (1.5). With the constraint $\sum_{j=1}^{n} \lambda_j = 1$ adjoined and setting $\theta^* = 1$ yields the following model,

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

The optimal slacks are $s_1^{-*} = 2$, $s_2^{-*} = s_1^{+*} = 0$, with $\lambda_3^* = 1$ and all other $\lambda_j^* = 0$.

1.4 Extensions to the CCR Model

A number of useful enhancements have appeared in the literature. Here, we limit our coverage to five of the extensions that illustrate the adaptability of the basic DEA methodology. The extensions discussed below allow an analyst to treat both nondiscretionary and categorical inputs and outputs and to incorporate judgment or ancillary managerial information. They are also easily extended to investigate efficiency changes over multiple time periods, and to measure congestion.

1.4.1 Nondiscretionary Inputs and Outputs

The above model formulations implicitly assume that all inputs and outputs are discretionary, i.e., can be controlled by the management of each DMU and varied at its discretion. Thus, failure of a DMU to produce maximal output levels with minimal input consumption results in a worsened efficiency score. However, there may exist exogenously fixed (or nondiscretionary) inputs or outputs that are beyond the control of a DMU's management. Instances from the DEA literature include snowfall or weather in evaluating the efficiency of maintenance units, soil characteristics and topography in different farms, number of competitors in the branches of a restaurant chain, local unemployment rates that affect the ability to attract recruits by different US Army recruitment stations, age of facilities in different universities, and number of transactions (for a purely gratis service) in library performance.

For example, Banker and Morey (1986a), whose formulations we use, illustrate the impact of exogenously determined inputs that are not controllable in an analysis of a network of fast-food restaurants. In their study, each of the 60 restaurants in the fast-food chain consumes six inputs to produce three outputs. The three outputs (all controllable) correspond to breakfast, lunch, and dinner sales. Only two of the six inputs, expenditures for supplies and expenditures for labor, are discretionary. The other four inputs (age of store, advertising level as determined by national headquarters, urban/rural location, and drive-in capability) are beyond the control of the individual restaurant manager in this chain.

The key to the proper mathematical treatment of a nondiscretionary variable lies in the observation that information about the extent to which a nondiscretionary input variable may be reduced is beyond the discretion of the individual DMU managers and thus cannot be used by them.

Suppose that the input and output variables may each be partitioned into subsets of discretionary (D) and nondiscretionary (N) variables. Thus,

$$I = \{1, 2, \ldots, m\} = I_{\mathrm{D}} \cup I_{\mathrm{N}} \text{ with } I_{\mathrm{D}} \cap I_{\mathrm{N}} = \emptyset$$

and

$$O = \{1, 2, \ldots, s\} = O_{\mathrm{D}} \cup O_{\mathrm{N}} \text{ with } O_{\mathrm{D}} \cap O_{\mathrm{N}} = \emptyset$$

where $I_{\mathrm{D}}, O_{\mathrm{D}}$ and $I_{\mathrm{N}}, O_{\mathrm{N}}$ refer to discretionary (D) and nondiscretionary (N) input, I, and output, O, variables, respectively, and ø is the empty set.

To evaluate managerial performance in a relevant fashion, we may need to distinguish between discretionary and nondiscretionary inputs as is done in the following modified version of a CCR model.

$$\begin{aligned}
&\min\ \theta - \varepsilon\left(\sum_{i\in I_{\mathrm{D}}} s_i^- + \sum_{r=1}^{s} s_r^+\right)\\
&\text{subject to}\\
&\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta x_{io} \qquad i \in I_{\mathrm{D}};\\
&\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = x_{io} \qquad i \in I_{\mathrm{N}}\\
&\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = y_{ro} \qquad r = 1, 2, \ldots, s;\\
&\lambda_j \geq 0 \qquad\qquad\qquad\qquad j = 1, 2, \ldots, n.
\end{aligned} \tag{1.11}$$

It is to be noted that the θ to be minimized appears only in the constraints for which $i \in I_{\mathrm{D}}$, whereas the constraints for which $i \in I_{\mathrm{N}}$ operate only indirectly

(as they should) because the input levels x_{io} for $i \in I_N$, are not subject to managerial control. It is also to be noted that the slack variables associated with I_N, the nondiscretionary inputs, are not included in the objective of (1.11) and hence the nonzero slacks for these inputs do not enter directly into the efficiency scores to which the objective is oriented.

The necessary modifications to incorporate nondiscretionary variables for the output-oriented CCR model is given by

$$\begin{aligned}
&\max \varphi + \varepsilon \left(\sum_{i=1}^{m} s_i^- + \sum_{r \in O_D} s_r^+ \right) \\
&\text{subject to} \\
&\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = x_{io} \qquad i = 1, 2, \ldots, m; \\
&\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = \varphi y_{ro} \qquad r \in O_D; \\
&\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = y_{ro} \qquad r \in O_N; \\
&\lambda_j \geq 0 \qquad\qquad\qquad\quad j = 1, 2, \ldots, n.
\end{aligned} \tag{1.12}$$

We should point out that there can be subtle issues associated with the concept of controllable outputs that may be obscured by the symmetry of the input/output model formulations. Specifically, switching from an input to an output orientation is not always as straightforward as it may appear. Interpretational difficulties for outputs not directly controllable may be involved as in the case of outputs influenced through associated input factors. An example of such an output would be sales that are influenced by advertising from the company headquarters, but are not directly controllable by district managers.

1.4.2 Categorical Inputs and Outputs

Our previous development assumed that all inputs and outputs were in the same category. However, this need not be the case as when some restaurants in a fast-food chain have a drive-in facility and some do not. See Banker and Morey (1986b) for a detailed discussion.

To see how this can be handled, suppose that an input variable can assume one of L levels (1, 2, …, L). These L values effectively partition the set of DMUs into categories. Specifically, the set of DMUs $K = \{1, 2, \ldots, n\} = K_1 \cup K_2 \cup \cdots \cup K_L$ where $K_f = \{j | j \in K$ and input value is $f\}$ and $K_i \cap K_j = \emptyset$, $i \neq j$. We wish to evaluate a DMU with respect to the envelopment surface determined for the units

contained in it and all preceding categories. The following model specification allows $\text{DMU}_o \in K_f$.

$$\begin{aligned}
&\min \theta \\
&\text{subject to} \\
&\sum_{j \in \bigcup_{f=1}^{K} K_f} x_{ij}\lambda_j + s_i^- = \theta x_{io} \quad i = 1, \ldots, m; \\
&\sum_{j \in \bigcup_{f=1}^{K} K_f} y_{rj}\lambda_j - s_r^+ = y_{ro} \quad r = 1, 2, \ldots, s; \\
&\lambda_j \geq 0 \qquad\qquad\qquad\qquad j = 1, 2, \ldots, n. \qquad (1.13)
\end{aligned}$$

Thus, the above specification allows one to evaluate all DMUs $l \in D_1$ with respect to the units in K_1, all DMUs $l \in K_2$ with respect to the units in $K_1 \cup K_2$, ..., all DMUs $l \in K_C$ with respect to the units in $\bigcup_{f=1}^{K_C} K_f$, etc. Although our presentation is for the input-oriented CCR model, it should be obvious that categorical variables can also be incorporated in this manner for any DEA model. In addition, the above formulation is easily implemented in the underlying LP solution algorithm via a candidate list.

The preceding development rests on the assumption that there is a natural nesting or hierarchy of the categories. Each DMU should be compared only with DMUs in its own and more disadvantaged categories, i.e., those operating under the same or worse conditions. If the categories are not comparable (e.g., public universities and private universities), then a separate analysis should be performed for each category.

1.4.3 Incorporating Judgment or A Priori Knowledge

A significant of the proposed extensions to DEA is the concept of restricting the possible range for the multipliers. In the CCR model, the only explicit restriction on multipliers is positivity, as noted for the $\varepsilon>0$ in (1.8). This flexibility is often presented as advantageous in applications of the DEA methodology, since a priori specification of the multipliers is not required and each DMU is evaluated in its best possible light.

In some situations, however, this complete flexibility may give rise to undesirable consequences, since it can allow a DMU to appear efficient in ways that are difficult to justify. Specifically, the model can assign unreasonably low or excessively high values to the multipliers in an attempt to drive the efficiency rating for a particular DMU as high as possible.

Three situations for which it has proven beneficial to impose various levels of control are the following:

1. The analysis would otherwise ignore additional information that cannot be directly incorporated into the model that is used, e.g., the envelopment model.
2. Management often has strong preferences about the relative importance of different factors and what determines best practice.
3. For a small sample of DMUs, the method fails to discriminate, and all may be efficient.

Using the multiplier models that duality theory makes available for introducing restrictions on the multipliers can affect solutions that can be obtained from the corresponding envelopment model. Proposed techniques for enforcing these additional restrictions include imposing upper and lower bounds on individual multipliers (Dyson and Thanassoulis 1988; Roll et al. 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).

To illustrate the general approach, suppose that we wish to incorporate additional inequality constraints of the following form into (1.3) or, more generally, into the multiplier model in Table 1.1:

$$\begin{aligned} \alpha_i \le \frac{v_i}{v_{i_o}} \le \beta_i, \quad & i = 1, \ldots, m \\ \delta_r \le \frac{\mu_r}{\mu_{r_o}} \le \gamma_r, \quad & r = 1, \ldots, s \end{aligned} \tag{1.14}$$

Here, v_{i_o} and μ_{r_o} represent multipliers that 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 $i = 1, \ldots, m$ and outputs $r = 1, \ldots, s$ 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. They may extend to "utils" or any other units that are regarded as pertinent. 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 to include bounds on these weights that better reflect local government preferences in measuring textile manufacturing performances.

Cook and Zhu (2008) developed a context-dependent AR approach where we can incorporate multiple sets of AR restrictions, with each reflecting the context for a particular subset of DMUs. The resulting modified DEA model, referred to as CAR-DEA, evaluates performance in a manner that more accurately captures the circumstances in which the DMUs operate.

There is another approach called the "cone-ratio envelopment approach" that can also be used for this purpose. See Charnes et al. (1990). We do not examine this approach in detail, but note only that the assurance region approach can also be given an interpretation in terms of cones. See Cooper et al. (1996).

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 are attained that appears to be reasonably satisfactory to decision-makers who cannot state the values for their preferences in an a priori manner.

The assurance region approach also greatly relaxes the conditions and widens the scope for use of a priori conditions. In some cases, the conditions to be comprehended may be too complex for explicit articulation, in which case additional possibilities are available from other recent advances. For instance, *instead* of imposing bounds on allowable variable values, the cone-ratio envelopment approach *transforms* the data. Brockett et al. (1997) provide an example in which a bank regulatory agency wanted to evaluate "risk coverage" as well as the "efficiency" of the banks under its jurisdiction. Bounds could not be provided on possible trade-offs between risk coverage and efficiency, so this was accomplished by using a set of banks identified as "excellent" (even when they were not members of the original (regulatory) set). Then, employing data from these excellent banks, a cone-ratio envelopment was used to transform the data into improved values that could be used to evaluate each of the regulated banks operating under widely varying conditions. This avoided the need for jointly specifying what was meant by "adequate" risk coverage and efficiency not only in each detail but also in all of the complex interplays between risk and efficiency that are possible in bank performances. The nonnegativity imposed on the slacks in standard CCR model was also relaxed. This then made it possible to identify deficiencies which were to be repaired by increasing expense items such as "bad loan allowances" (as needed for risk coverage) even though this worsened efficiency as evaluated by the transformed data.

There are in fact a number of ways in which we can incorporate a priori knowledge or requirements as conditions into DEA models. For examples see the "Prioritization Models" of Cook et al. (1993) and the "Preference Structure Model" of Zhu (1996b). See Chap. 4 for choices and uses of DEA weights (multipliers).

1.4.4 Window Analysis

In the examples of the previous sections, each DMU was observed only once, i.e., each example was a cross-sectional analysis of data. In actual studies, observations for DMUs are frequently available over multiple time periods (time series data), and it is often important to perform an analysis where interest focuses on changes in efficiency over time. In such a setting, it is possible to perform DEA over time by

using a moving average analogue, where a DMU in each different period is treated as if it were a "different" DMU. Specifically, a DMU's performance in a particular period is contrasted with its performance in other periods in addition to the performance of the other DMUs.

The *window analysis* technique that operationalizes the above procedure can be illustrated with the study of aircraft maintenance operations, as described in Charnes et al. (1985). In this study, data were obtained for 14 ($n = 14$) tactical fighter wings in the US Air Force over seven ($p = 7$) monthly periods. To perform the analysis using a 3-month ($w = 3$) window, one proceeds as follows.

Each DMU is represented as if it were a different DMU for each of the three successive months in the first window (Ml, M2, M3) consisting of the months at the top of Table 1.4. An analysis of the 42 ($= nw = 3 \times 14$) DMUs can then be performed. The window is then shifted one period by replacing M1 with M4, and an analysis is performed on the second three-month set (M2, M3, M4) of these 42 DMUs. The process continues in this manner, shifting the window forward one period each time and concluding with the final (fifth) analysis of 42 DMUs for the last three months (M5, M6, M7). (In general, one performs $p - w + 1$ separate analyses, where each analysis examines nw DMUs.)

Table 1.4 illustrates the results of this analysis in the form of efficiency scores for the performance of the airforce wings as taken from Charnes et al. (1985). The structure of this table portrays the underlying framework of the analysis. For the first "window," wing A is represented in the constraints of the DEA model as though it were a different DMU in months 1, 2, and 3. Hence, when wing 1 is evaluated for its month-1 efficiency, its own performance data for months 2 and 3 are included in the constraint sets along with similar performance data of the other wings for months 1, 2, and 3. Thus, the results of the "first window" analysis consist of the 42 scores under the column headings for month 1 to month 3 in the first row for each wing. For example, wing A had efficiency ratings of 97.89, 97.31, and 98.14 for its performance in months 1, 2, and 3, respectively, as shown in the first row for Wing A in Table 1.4. The second row of data for each wing is the result of analyzing the second window of 42 DMUs, which result from dropping the month-1 data and appending the month-4 data.

The arrangement of the results of a window analysis as given in Table 1.4 facilitates the identification of trends in performance, the stability of reference sets, and other possible insights. For illustration, "row views" clarify performance trends for wings E and M. Wing E improved its performance in month 5 relative to prior performance in months 3 and 4 in the third window, while wing M's performance appears to deteriorate in months 6 and 7. Similar "column views" allow comparison of wings (DMUs) across different reference sets and hence provide information on the stability of these scores as the reference sets change.

The utility of Table 1.4 can be further extended by appending columns of summary statistics (mean, median, variance, range, etc.) for each wing to reveal the relative stability of each wing's results. See, for instance, the drop in the efficiency from 93.74 to 82.54 in month 4 for Wing K.

The window analysis technique represents one area for further research to extend DEA. For example, the problem of choosing the width for a window (and

Table 1.4 Window analysis with 3-month window

Wing	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7
Wing-A	97.89	97.31	98.14				
		97.36	97.53	97.04			
			96.21	95.92	94.54		
				95.79	94.63	97.64	
					94.33	97.24	97.24
Wing-B	93.90	95.67	96.14				
		96.72	96.42	94.63			
			95.75	94.14	93.26		
				94.54	93.46	96.02	
					93.02	96.02	94.49
Wing-C	93.77	91.53	95.26				
		91.77	95.55	94.29			
			93.21	95.04	94.83		
				93.20	93.09	92.21	
					93.59	92.32	92.83
Wing-D	99.72	96.15	95.06				
		97.91	95.70	100.0			
			94.79	100.0	94.51		
				99.71	94.39	94.76	
					94.95	94.67	89.37
Wing-E	100.0	100.0	100.0				
		100.0	100.0	100.0			
			98.97	99.05	100.0		
				99.37	100.0	100.0	
					100.0	100.0	100.0
Wing-F	97.42	93.48	96.07				
		93.60	96.24	93.56			
			94.46	91.75	92.49		
				91.73	92.32	92.35	
					92.68	91.98	99.64
Wing-G	90.98	92.80	95.96				
		93.67	96.80	99.52			
			93.34	94.48	91.73		
				91.94	89.79	95.58	
					89.35	95.14	96.38
Wing-H	100.0	100.0	100.0				
		100.0	100.0	100.0			
			100.0	100.0	100.0		
				100.0	100.0	100.0	
					100.0	100.0	100.0
Wing-I	99.11	95.94	99.76				
		96.04	100.0	100.0			
			98.16	98.99	94.59		
				98.97	94.62	99.16	
					94.68	98.92	97.28

(continued)

Table 1.4 (continued)

Wing	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7
Wing-J	92.85	90.90	91.62				
		91.50	92.12	94.75			
			90.26	93.39	93.83		
				92.92	93.84	95.33	
					94.52	96.07	94.43
Wing-K	86.25	84.42	84.03				
		84.98	84.47	93.74			
			83.37	82.54	80.26		
				82.39	80.14	79.58	
					80.96	78.66	79.75
Wing-L	100.0	100.0	100.0				
		100.0	100.0	99.55			
			100.0	99.39	97.39		
				100.0	96.85	100.0	
					96.66	100.0	100.0
Wing-M	100.0	100.0	100.0				
		100.0	100.0	100.0			
			100.0	100.0	100.0		
				100.0	100.0	98.75	
					100.0	98.51	99.59
Wing-N	100.0	100.0	98.63				
		100.0	100.0	100.0			
			99.45	100.0	100.0		
				100.0	100.0	100.0	
					100.0	100.0	100.0

the sensitivity of DEA solutions to window width) is currently determined by trial and error. Similarly, the theoretical implications of representing each DMU as if it were a different DMU for each period in the window remain to be worked out in full detail.

1.5 Allocative and Overall Efficiency

To this point, we have confined attention to "technical efficiency" which, as explained immediately after definition 1.2, does not require a use of prices or other "weights." Now, we extend the analysis to situations in which unit prices and unit costs are available. This allows us to introduce the concepts of "allocative" and "overall" efficiency and relate them to "technical efficiency" in a manner first introduced by Farrell.

For this introduction, we utilize Fig. 1.5 in which the solid line segments connecting points ABCD constitute an "isoquant" or "level line' that represents

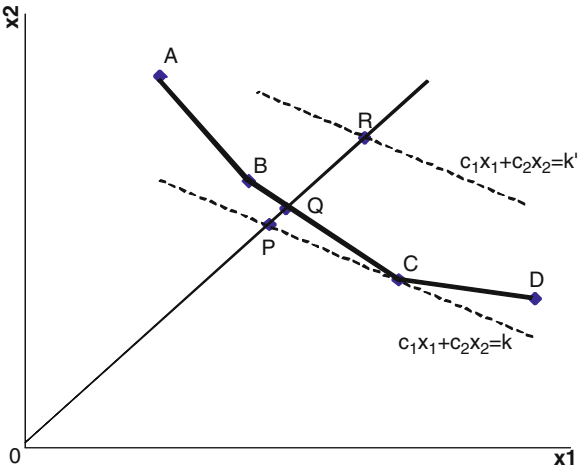


Fig. 1.5 Allocative and overall efficiency

the different amounts of two inputs (x_1, x_2), which can be used to produce the same amount (usually one unit) of a given output. This line represents the "efficiency frontier" of the "production possibility set" because it is not possible to reduce the value of one of the inputs without increasing the other input if one is to stay on this isoquant.

The dashed line represents an isocost (=budget) line for which the (x_1, x_2) pairs on this line yield the same total cost, when the unit costs are c_1 and c_2, respectively. When positioned on C the total cost is k. However, shifting this budget line upward in parallel fashion until it reaches a point of intersection with R would increase the cost to $k'>k$. In fact, as this figure shows, k is the minimum total cost needed to produce the specified output since any parallel shift downward below C would yield a line that fails to intersect the production possibility set. Thus, the intersection at C gives an input pair (x_1, x_2) that minimizes the total cost of producing the specified output amount, and the point C is, therefore, said to be "allocatively" as well as "technically" efficient.

Now, let R represent an observation that produced this same output amount. The ratio $0 \leq \mathrm{OQ}/\mathrm{OR} \leq 1$ is said to provide a "radial" measure of technical efficiency, with $0 \leq 1 - (\mathrm{OQ}/\mathrm{OR}) \leq 1$ yielding a measure of technical inefficiency. Actually, this "radial measure" is a ratio of two measures of distance, one of which measures the distance from the origin to Q and the other which measures the distance from the origin to R to obtain $0 \leq d(\mathrm{O},\mathrm{Q})/d(\mathrm{O},\mathrm{R}) \leq 1$ where $d(\ldots)$ means "distance." See Bardhan et al. (1996) for a proof in terms of the Euclidean measure of distance – although other measures of distance may also be used.

Now consider the point P which is at the intersection of this cost line through C with the ray from the origin to R. We can also obtain a radial measure of "overall

efficiency" from the ratio $0 \leq \mathrm{OP}/\mathrm{OR} \leq 1$. In addition, we can form the ratio $0 \leq \mathrm{OP}/\mathrm{OQ} \leq 1$ to obtain a measure of what Farrell referred to as "price efficiency" but is now more commonly called "allocative efficiency." Finally, we can relate these three measures to each other by noticing that

$$\frac{\mathrm{OP}}{\mathrm{OQ}}\frac{\mathrm{OQ}}{\mathrm{OR}} = \frac{\mathrm{OP}}{\mathrm{OR}} \tag{1.15}$$

which we can verbalize by saying that the product of allocative and technical efficiency eqauls overall efficiency in these radial measures.

To implement these ideas we use the following model, as taken from Cooper et al. (2007, p. 236),

$$\begin{aligned}
&\min \sum_{i=1}^{m} c_{io} x_i \\
&\text{subject to} \\
&\sum_{j=1}^{n} x_{ij}\lambda_j \leq x_i, \qquad i = 1,\ldots,m \\
&\sum_{j=1}^{n} y_{rj}\lambda_j \geq y_{ro}, \qquad r = 1,\ldots,s \\
&L \leq \sum_{j=1}^{n} \lambda_j \leq U
\end{aligned} \tag{1.16}$$

where the objective is to choose the x_i and λ_j values to minimize the total cost of satisfying the output constraints. The c_{io} in the objective represent unit costs. This formulation differs from standard models, as in Färe et al. (1985, 1994), in that these unit costs are allowed to vary from one DMU_o to another in (1.16). In addition the values of $\sum_{j=1}^{n} \lambda_j$ are limited above by U and below by L – according to the returns-to-scale conditions that are imposed. See next chapter. Here we only note that the choice $L = U = 1$ makes this a BCC model, whereas $L = 0$, $U = \infty$ converts it to a CCR model. See the discussion following Table 1.1. Finally, using the standard approach, we can obtain a measure of relative cost (=overall) efficiency by utilizing the ratio

$$0 \leq \frac{\sum_{i=1}^{m} c_{io} x_i^*}{\sum_{i=1}^{m} c_{io} x_{io}} \leq 1 \tag{1.17}$$

where the x_i^* are the optimal values obtained from (1.16) and the x_{io} are the observed values for DMU$_o$.

The use of a ratio like (1.17) is standard and yields an easily understood measure. It has shortcomings, however, as witness the following example from Tone and Sahoo (2003): Let γ_a and γ_b represent cost efficiency, as determined from (1.17), for DMUs a and b. Now suppose $x_{ia}^* = x_{ib}^*$ and $x_{ia} = x_{ib}, \forall i$, but $c_{ia} = 2c_{ib}$ so that the unit costs for a are twice as high as for b in every input. We then have

$$\gamma_a = \frac{\sum_{i=1}^{m} c_{ia}x_{ia}^*}{\sum_{i=1}^{m} c_{ia}x_{ia}} = \frac{\sum_{i=1}^{m} 2c_{ib}x_{ib}^*}{\sum_{i=1}^{m} 2c_{ib}x_{ib}} = \frac{\sum_{i=1}^{m} c_{ib}x_{ib}^*}{\sum_{i=1}^{m} c_{ib}x_{ib}} = \gamma_b \tag{1.18}$$

Thus, as might be expected with a use of ratios, important information may be lost since $\gamma_a = \gamma_b$ conceals the fact that a is twice as costly as b.

1.6 Profit Efficiency

We now introduce another type of model called the "additive model" to evaluate technical inefficiency. First introduced in Charnes et al. (1985) this model has the form

$$\begin{aligned}
&\max \sum_{r=1}^{s} s_r^+ + \sum_{i=1}^{m} s_i^- \\
&\text{subject to} \\
&y_{ro} = \sum_{\mathrm{j}=1}^{n} y_{rj}\lambda_j - s_r^+, \quad r = 1, 2, \ldots, s \\
&x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-, \quad i = 1, 2, \ldots, m \\
&1 = \sum_{j=1}^{n} \lambda_j \\
&0 \le \lambda_j, s_r^+ s_i^-; \quad \forall i, j, r.
\end{aligned} \tag{1.19}$$

This model uses a metric that differs from the one used in the "radial measure" model which uses what is called the "ℓ_1 metric" in mathematics, and the "city block metric" in operations research. See Appendix A in Charnes and Cooper (1961). It also dispenses with the need for distinguishing between an "output" and an "input" orientation as was done in the discussion leading up to (1.10) because the objective in (1.19) simultaneously maximizes outputs and minimizes inputs – in the sense of

vector optimizations. This can be seen by utilizing the solution to (1.19) to introduce new variables $\hat{y}_{ro}, \hat{x}_{io}$ defined as follows:

$$\begin{aligned} \hat{y}_{ro} &= y_{ro} + s_r^{+*} \geq y_{ro}, \quad r = 1, \ldots, s, \\ \hat{x}_{io} &= x_{io} - s_i^{-*} \leq x_{io}, \quad i = 1, 2, \ldots, m. \end{aligned} \tag{1.20}$$

Now, note that the slacks are all independent of each other. Hence, an optimum is not reached until it is not possible to increase an output $\hat{y}_{ro}$ or reduce an input $\hat{x}_{io}$ without decreasing some other output or increasing some other input. The following theorem, which follows immediately, is proved in Cooper et al. (2007),

Theorem 1.2. DMU$_o$ is efficient if and only if all slacks are zero in an optimum solution.

As proved in Ahn et al. (1988), one can also relate solutions in the additive model to those in the radial measure model via

Theorem 1.3. DMU$_o$ is efficient for an additive model if and only if it is efficient for the corresponding radial model.

Here, the term "corresponding" means that the constraint sets are the same so that $\sum_{j=1}^{n} \lambda_j = 1$ appears as a constraint in the additive model if and only if it also appears in the radial measure model to which it is being compared.

We now use the class of additive models to develop a different route to treating technical, allocative and overall inefficiencies and their relations to each other. This can help to avoid difficulties in treating possibilities such as "negative" or "zero" profits, which are not easily treated by the ratio approaches, as in (1.17), which are commonly used in the DEA literature. See the discussion in the appendix to Cooper et al. (1999a, b) from which the following development is taken. See also Chap. 8 in Cooper et al. (2007).

First, we observe that we can multiply the output slacks by unit prices and the input slacks by unit costs after we have solved (1.19) and thereby accord a monetary value to this solution. Then, we can utilize (1.20) to write

$$\begin{aligned} \sum_{r=1}^{s} p_{ro} s_r^{+*} + \sum_{i=1}^{m} c_{io} s_i^{-*} &= \left(\sum_{r=1}^{s} p_{ro} \hat{y}_{ro} - \sum_{i=1}^{m} p_{ro} y_{ro} \right) + \left(\sum_{i=1}^{m} c_{io} x_{io} - \sum_{i=1}^{m} c_{io} \hat{x}_{io} \right) \\ &= \left(\sum_{r=1}^{s} p_{ro} \hat{y}_{ro} - \sum_{i=1}^{m} c_{io} \hat{x}_{io} \right) - \left(\sum_{r=1}^{s} p_{ro} y_{ro} - \sum_{i=1}^{m} c_{io} x_{io} \right). \end{aligned} \tag{1.21}$$

From the last pair of parenthesized expressions we find that, at an optimum, the objective in (1.19) after multiplication by unit prices and costs is equal to the profit available when production is technically efficient minus the profit obtained from the

observed performance. Hence, when multiplied by unit prices and costs, the solution to (1.19) provides a measure in the form of the amount of the profits lost by not performing in a technically efficient manner term by term if desired.

Remark. We can, if we wish restate this measure in ratio form because, by definition,

$$\sum_{r=1}^{s} p_{ro}\hat{y}_{ro} - \sum_{i=1}^{m} c_{io}\hat{x}_{io} \geq \sum_{r=1}^{s} p_{ro}y_{ro} - \sum_{i=1}^{m} c_{io}x_{io}$$

Therefore,

$$1 \geq \frac{\sum_{r=1}^{s} p_{ro}y_{ro} - \sum_{i=1}^{m} c_{io}x_{io}}{\sum_{r=1}^{s} p_{ro}\hat{y}_{ro} - \sum_{i=1}^{m} c_{io}\hat{x}_{io}} \geq 0$$

and the upper bound is attained if and only if performance is efficient.

We can similarly, develop a measure of allocative efficiency by means of the following additive model:

$$\begin{aligned}
&\max \sum_{r=1}^{s} p_{ro}\hat{s}_r^{+} + \sum_{i=1}^{m} c_{io}\hat{s}_i^{-} \\
&\text{subject to} \\
&\hat{y}_{ro} = \sum_{j=1}^{n} y_{rj}\hat{\lambda}_j - \hat{s}_r^{+}, \quad r = 1, 2, \ldots, s \\
&\hat{x}_{io} = \sum_{j=1}^{n} x_{ij}\hat{\lambda}_j - \hat{s}_i^{-}, \quad i = 1, 2, \ldots, m \\
&1 = \sum_{j=1}^{n} \hat{\lambda}_j \\
&0 \leq \hat{\lambda}_j \forall j; \hat{s}_i^{-}, \ \hat{s}_r^{+} \text{ free} \quad \forall i, r.
\end{aligned} \tag{1.22}$$

Comparison with (1.19) reveals the following differences: (1) the objective in (1.19) is replaced by one which is monetized (2) the y_{ro} and x_{io} in (1.19) are replaced by $\hat{y}_{ro}$ and $\hat{x}_{io}$ in (1.22) as obtained from (1.20) and, finally, (3) the slack values in (1.22) are not constrained in sign as is the case in (1.19). This last relaxation, we might note, is needed to allow for substitutions between the different output and the different input amounts, as may be needed to achieve the proportions required for allocative efficiency. See Cooper et al. (2000 b).

Finally, we use the following additive model to evaluate overall (profit) efficiency – called "graph efficiency" in Färe et al. (1985, 1994) – see also Färe and Grosskopf (1996)

Table 1.5 Price–cost–profit data

	DMU_1	DMU_2	DMU_3	$
y	4	4	2	6
x_1	4	2	4	4
x_2	2	4	6	2
π	4	8	−16	

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

subject to

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

$$x_{io} = \sum_{j=1}^{n} x_{ij} \lambda_j - \hat{s}_i^-, \quad i = 1, 2, \ldots, m$$

$$1 = \sum_{j=1}^{n} \lambda_j$$

$$0 \leq \lambda_j \forall j; s_i^-, \; s_r^+ \text{ free } \forall i, r. \tag{1.23}$$

We then have the relation set forth in the following:

Theorem 1.4. The value (=total profit foregone) of overall inefficiency for DMU_o as obtained from (1.23) is equal to the value of technical inefficiency as obtained from (1.19) plus the value of allocative inefficiency as obtained from (1.22). i.e.,

$$\max \left(\sum_{r=1}^{s} p_{ro} s_r^+ + \sum_{i=1}^{m} c_{io} s_i^- \right) = \left(\sum_{r=1}^{s} p_{ro} s_r^{+*} + \sum_{i=1}^{m} c_{io} s_i^{-*} \right) + \left(\sum_{r=1}^{s} p_{ro} \hat{s}_r^{+*} + \sum_{i=1}^{m} c_{io} \hat{s}_i^{-*} \right).$$

Table 1.5, as adapted from Cooper et al. (2007), can be used to construct an example to illustrate this theorem. The body of the table records the performances of 3 DMUs in terms of the amount of the one output they all produce, y, and the amount of the two inputs x_1, x_2 they all use. For simplicity, it is assumed that they all receive the same unit price $p = 6$ as recorded on the right and incur the same unit costs $c_1 = \$4$ and $c_2 = \$2$ for the inputs as shown in the rows with which they are associated. The bottom of each column records the profit, π, made by each DMU in the periods of interest.

DMU_3, as shown at the bottom of its column, is, by far, the worst performer having experienced a loss of \$16. This loss, however, does not account for all of the lost profit possibilities. To discover this value, we turn to (1.23) and apply it to the

data in Table 1.5. This produces the following model to evaluate the overall inefficiency of DMU_3.

$$\begin{aligned}
&\max 6s^+ + 4s_1^- + 2s_2^- \\
&\text{subject to} \\
&\quad 2 = 4\lambda_1 + 4\lambda_2 + 2\lambda_3 - s^+ \\
&\quad 4 = 4\lambda_1 + 2\lambda_2 + 4\lambda_3 \quad + s_1^- \\
&\quad 6 = 2\lambda_1 + 4\lambda_2 + 6\lambda_3 \quad - s_2^- \\
&\quad 1 = \lambda_1 + \lambda_2 + \lambda_3 \\
&\quad 0 \le \lambda_1, \lambda_2, \lambda_3; s^+, s_1^-, s_2^- \text{ free.}
\end{aligned} \tag{1.24}$$

An optimum solution to this problem is $\lambda_2^* = 1$, $s^{+*} = 2$, $s_1^{-*} = 2$, $s_2^{-*} = 2$ and all after variables zero. Utilizing the unit price and costs exhibited in the objective of (1.24), we, therefore, find

$$6s^{+*} + 4s_1^{-*} + 2s_1^{-*} = \$6 \times 2 + \$4 \times 2 + \$2 \times 2 = \$24.$$

This is the value of the foregone profits arising from inefficiencies in the performance of DMU_3. Eliminating these inefficiencies would have wiped out the \$16 loss and replaced it with an \$8 profit. This is the same amount of profit as DMU_2, which is the efficient performer used to effect this evaluation of DMU_3 via $\lambda_2^* = 1$ in the above solution.

We now utilize theorem 1.4 to further identify the sources of this lost profit. For this purpose, we first apply (1.19) to the data of Table 1.5 to determine the lost profits from the technical inefficiency of DMU_3 via

$$\begin{aligned}
&\max s^+ + s_1^- + s_1^- \\
&\text{subject to} \\
&\quad 2 = 4\lambda_1 + 4\lambda_2 + 2\lambda_3 - s^+ \\
&\quad 4 = 4\lambda_1 + 2\lambda_2 + 4\lambda_3 \quad + s_1^- \\
&\quad 6 = 2\lambda_1 + 4\lambda_2 + 6\lambda_3 \quad - s_2^- \\
&\quad 1 = \lambda_1 + \lambda_2 + \lambda_3 \\
&\quad 0 \le \lambda_1, \lambda_2, \lambda_3, s^+, s_1^-, s_2^-.
\end{aligned} \tag{1.25}$$

This has an optimum with $\lambda_1^* = 1$, $s^{+*} = 2$, $s_1^{-*} = 0$, $s_2^{-*} = 4$ and all other variables zero so that multiplying these values by their unit price and unit costs we find that the lost profits due to technical inefficiencies are

$$\$6 \times 2 + \$4 \times 0 + \$2 \times 4 = \$20.$$

For allocative inefficiency, we apply (1.20) and (1.22) to the data in Table 1.5 and get.

Table 1.6 Solution detail

Model variable	Overall	Technical	Allocative
s^+	2	2	0
s_1^-	2	0	2
s_2^-	2	4	−2
π	24	20	4

$$
\begin{aligned}
&\max 6\hat{s}^+ + 4\hat{s}_1^- + 2\hat{s}_2^- \\
&\text{subject to} \\
&\quad 4 = 4\hat{\lambda}_1 + 4\hat{\lambda}_2 + 2\hat{\lambda}_3 - \hat{s}^+ \\
&\quad 4 = 4\hat{\lambda}_1 + 2\hat{\lambda}_2 + 4\hat{\lambda}_3 \qquad + \hat{s}_1^- \\
&\quad 2 = 2\hat{\lambda}_1 + 4\hat{\lambda}_2 + 6\hat{\lambda}_3 \qquad\quad + \hat{s}_2^- \\
&\quad 1 = \hat{\lambda}_1 + \hat{\lambda}_2 + \hat{\lambda}_3 \\
&\quad 0 \le \hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3; \hat{s}^+, \hat{s}_1^-, \hat{s}_2^- \text{ free.}
\end{aligned}
$$

An optimum is $\hat{\lambda}_2^* = 1, \hat{s}^{+*} = 0, \hat{s}_1^{-*} = 2, \hat{s}_2^{-*} = -2$ with all other variables zero, so the profit lost from allocative efficiency is

$$\$4 \times 2 + \$2(-2) = \$4,$$

which accounts for the remaining \$4 of the \$24 lost profit obtained from overall inefficiency via (1.24). Here, we might note that an increase in $\hat{x}_2$ – see the second expression in (1.20) – is more than compensated for by the offsetting decrease in $\hat{x}_1$ en route to the proportions needed to achieve allocative efficiency.

Finally, we supply a tabulation obtained from these solutions in Table 1.6.

Remark 1. Adding the figures in each row of the last two columns yields the corresponding value in the column under overall efficiency. This will always be true for the dollar value in the final row, by virtue of theorem 1.3, but it need not be true for the other rows because of the possible presence of alternate optima.

Remark 2. The solutions need not be "units invariant." That is, the optimum solutions for the above models may differ if the unit prices, and unit costs used are stated in different units. See pp. 228 ff. in Cooper et al. (2000a, b) for a more detailed discussion and methods for making the solutions units invariant.

1.7 Recent Developments

In a recent paper by Cook and Seiford (2009), a number of new DEA developments are presented with respect the modeling of DMU internal structures and to the status of inputs and outputs.

The network DEA model originated by Färe and Grosskopf (1996) is built around the concept of subtechnologies within the "black box" of DEA. This approach allows one to examine in more detail the inner workings of the production process.

In a different line of work, several DEA-based approaches have been used to examine buyer–supplier supply chain (or two-stage process) settings. The important issue is that of deriving a measure of overall efficiency as opposed to looking only at the efficiencies of individual stages of the processes or supply chains. Chen and Zhu (2004) provide two approaches in modeling two-stage processes. Zhu (2009) presents a DEA-based supply chain model to both measure the overall efficiency, and that of its members.

A number of supply chain approaches due to Liang et al. (2006, 2008) are built on game theoretic constructs. Their two principal models are (1) a noncooperative model and (2) a cooperative model. In the noncooperative model, they view the seller as the leader and buyer as the follower. In the first stage, one optimizes the leader's efficiency score and then maximizes (in the second stage) that of the follower, with the constraint that the multipliers used must be such that the first stage (leader) score remains unchanged. The resulting model is a nonlinear parametric programming problem. In the cooperative game model, no leader–follower assumption is made. Cook et al. (2010) review the existing DEA approaches for dealing with two-stage processes.

Other related DEA developments include multicomponent/parallel models of Cook et al. (2000) and hierarchical/nested models of Cook et al. (1998) and Cook and Green (2005).

In addition to the treatment of nondiscretionary and categorical variables, ordinal or rank order data are modeled in DEA approach. The original DEA models incorporating rank order variables are due to Cook et al. (1993, 1996). Cooper et al. (1999a, b) examined the DEA structure in the presence of what they termed imprecise data (IDEA). Zhu (2003b) and others have extended the Cooper et al. (1999a, b) model. While various forms of imprecision are looked at under the umbrella of IDEA, the principal focus is on rank order data. In a paper by Cook and Zhu (2006), rank order variables and IDEA are revisited, and both discrete and continuous projection models are discussed. It is shown that the IDEA approach for rank data is equivalent to the Cook et al. (1993, 1996) methodology. Chapter 6 discusses the treatment of qualitative data in DEA.

The usual variables in DEA are such that more is better for outputs, and less is better for inputs. In some situations, however, a factor can behave opposite to this; consider, for example, air pollution as one of the outputs from power plants. A number of authors have addressed this issue (In particular, Scheel (2001), Seiford and Zhu (2002), Färe and Grosskopf (2004), and Hua and Bin (2007)). Approaches range from linear transformations of the original data to the use of directional distance functions.

1.8 Conclusions

This chapter provides an introduction to DEA and some of its uses. However, it is far from having exhausted the possibilities that DEA offers. For instance, we focus here on what is referred to in the DEA (and economics) literature as technical efficiency. For perspective, we conclude with discussions of allocative and overall efficiency when costs or profits are of interest and the data are available. This does not exhaust the possibilities. There are still other types of efficiency that can be addressed with DEA. For instance, returns-to-scale inefficiencies, as covered in the next chapter, can offer additional possibilities that identify where additional shortcomings can appear when unit prices or unit costs are available, which are of interest to potential users. We have not even fully exploited uses of our technical inefficiency models, as developed in this chapter. For instance, uses of DEA identify DMUs that enter into the optimal evaluations and hence can serve as "benchmark DMUs" to determine how best to eliminate these inefficiencies.

Topics such as these are discussed and extended to probabilistic and other characterizations in some of the chapters that follow. However, the concept of technical inefficiency provides a needed start, which will turn out to be basic for all of the other types of efficiency that may be of interest. Technical efficiency is also the most elemental of the various efficiencies that might be considered in that it requires only minimal information and minimal assumptions for its use. It is also fundamental because other types of efficiency such as allocative efficiency and returns to scale efficiency require technical efficiency to be attained before these can be achieved. All these are treated in the chapters that follow.

References

Afriat S. Efficiency estimation of production functions. Int Econ Rev. 1972;13:568–98.

Ahn T, Charnes A, Cooper WW. Efficiency characterizations in different DEA models. Soc Econ Plann Sci. 1988;22:253–7.

Arnold V, Bardhan I, Cooper WW, Gallegos A. Primal and dual optimality in computer codes using two-stage solution procedures in DEA. In: Aronson J, Zionts S, editors. Operations research methods, models and applications. Westpost, CT: Quorum Books; 1998.

Banker R, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manag Sci. 1984;30:1078–92.

Banker RD, Morey RC. Efficiency analysis for exogenously fixed inputs and outputs. Oper Res. 1986a;34(4):513–21.

Banker RD, Morey RC. The use of categorical variables in data envelopment analysis. Manag Sci. 1986b;32(12):1613–27.

Bardhan I, Bowlin WF, Cooper WW, Sueyoshi T. Models and measures for efficiency dominance in DEA, part I: additive models and MED measures. J Oper Res Soc Jpn. 1996;39:322–32.

Bessent A, Bessent W, Charnes A, Cooper WW, and Thorogood (1983). Evaluation of Educational Proposals by means of DEA, Educational Administrative Quarterly 14, 82–107. Reprinted in Management Science in Higher Education: Methods and Studies (New York: Elsevier Publishing co., 1987).

Brockett PL, Charnes A, Cooper WW, Huang ZM, Sun DB. Data transformations in DEA cone ratio envelopment approaches for monitoring bank performances. Eur J Oper Res. 1997;98 (2):250–68.

Bulla S, Cooper WW, Parks KS, Wilson D. Evluating efficiencies in Turbo-Fan jet engines in multiple inputs–outputs context approaches. Propul Power. 2000;16:431–9.

Charnes A, Cooper WW. Management models and industrial applications of linear programming. New York, NY: Wiley; 1961.

Charnes A, Cooper WW. Programming with linear fractional functionals. Nav Res Logist Q. 1962;9:181–5.

Charnes A, Cooper WW, Rhodes E. 1978, Measuring the efficiency of decision making units, European Journal of Operational Research 2, 429–444. Also, 1979, Short Communication, European Journal of Operational Research 3, 339–340.

Charnes A, Clarke CT, Cooper WW, Golany B. A developmental study of data envelopment analysis in measuring the efficiency of maintenance units in the US air forces. Annals of operation research. 1985. Vol. 2 p. 95–112.

Charnes A, Cooper WW, Sun DB, Huang ZM. Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks. J Econ. 1990;46:73–91.

Charnes A, Cooper WW, Wei QL, Huang ZM. Cone ratio data envelopment analysis and multi-objective programming. Int J Syst Sci. 1989;20:1099–118.

Charnes A, Cooper WW, Golany B, Seiford L, Stutz J. Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. J Econ. 1985b;30:91–l07.

Chen Y, Zhu J. Measuring information technology's indirect impact on firm performance. Inform Technol Manag J. 2004;5(1–2):9–22.

Cook WD, Seiford LM. Data envelopment analysis (DEA) – Thirty years on. Eur J Oper Res. 2009;192:1–17.

Cook WD, Kress M, Seiford LM. On the use of ordinal data in data envelopment analysis. J Oper Res Soc. 1993;44:133–40.

Cook WD, Kress M, Seiford LM. Data envelopment analysis in the presence of both quantitative and qualitative factors. J Oper Res Soc. 1996;47:945–53.

Cook WD, Zhu J. Context-dependent assurance regions in DEA. Oper Res. 2008;56(1):69–78.

Cook WD, Liang L, Zhu J. Measuring performance of two-stage network structures by DEA: a review and future perspective. Omega. 2010;38:423–30.

Cook WD, Hababou M, Tuenter H. Multi-component efficiency measurement and shared inputs in data envelopment analysis: an application to sales and service performance in bank branches. J Product Anal. 2000;14:209–24.

Cook WD, Green RH. Evaluating power plant efficiency: a hierarchical model. Comput Oper Res. 2005;32:813–23.

Cook WD, Chai D, Doyle J, Green RH. Hierarchies and groups in DEA. J Product Anal. 1998;10:177–98.

Cook WD, Zhu J. Rank order data in DEA: a general framework. Eur J Oper Res. 2006;174 (2):1021–38.

Cooper WW, Thompson RG, Thrall RM. Extensions and new developments in data envelopment analysis. Ann Oper Res. 1996;66:3–45.

Cooper WW, Seiford LM, Tone 2nd K, editors. Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software. Boston: Kluwer; 2007.

Cooper WW, Seiford LM, Zhu J. A unified additive model approach for evaluating inefficiency and congestion with associated measures in DEA. Soc Econ Plann Sci. 2000a;34(1):1–25.

Cooper WW, Park KS, Pastor JT. Marginal rates and Elasticities of substitution in DEA. J Product Anal. 2000b;13(2000):105–23.
Cooper WW, Park KS, Pastor JT. RAM: a range adjusted measure of inefficiency for use with additive models and relations to other models and measures in DEA. J Product Anal. 1999a;11:5–42.
Cooper WW, Park KS, Yu G. IDEA and AR-IDEA: models for dealing with imprecise data in DEA. Manag Sci. 1999b;45:597–607.
Debreu G. The coefficient of resource utilization. Econometrica. 1951;19:273–92.
Dyson RG, Thanassoulis E. Reducing weight flexibility in data envelopment analysis. J Oper Res Soc. 1988;39(6):563–76.
Emrouznejad A, Parker BR, Tavares G. Evaluation of research in efficiency and productivity: a survey and analysis of the first 30 years of scholarly literature in DEA. Soc Econ Plann Sci. 2008;42:151–7.
Färe R, Grosskopf S, Lovell CAK. The measurement of efficiency of production. Boston: Kluwer Nijhoff Publishing Co.; 1985.
Färe R, Grosskopf S, Lovell CAK. Production frontiers. Cambridge: Cambridge University Press; 1994.
Färe R, Grosskopf S. Intertemporal production frontiers: with dynamic DEA. Boston, MA: Kluwer Academic; 1996.
Färe R, Grosskopf S. Modelling undesirable factors in efficiency evaluation: Comment. Eur J Oper Res. 2004;157:242–5.
Farrell MJ. The measurement of productive efficiency. J Roy Stat Soc A. 1957;120:253–81.
Hua Z, Bin Y. DEA with undesirable factors. (Chapter 6). In: Zhu J, Cook WD, editors. Modeling data irregularities and structural complexities in data envelopment analysis. Boston: Springer Science; 2007.
Koopmans TC, editor. Analysis of production as an efficient combination of activities. New York: Wiley; 1951.
Liang LF, Yang F, Cook WD, Zhu J. DEA models for supply chain efficiency evaluation. Ann Oper Res. 2006;145(1):35–49.
Liang L, Cook WD, Zhu J. DEA Models for two-stage processes: game approach and efficiency decomposition. Nav Res Logist. 2008;55:643–53.
Roll Y, Cook WD, Golany B. Controlling factor weights in data envelopment analysis. IIE Trans. 1991;23:2–9.
Scheel H. Undesirable outputs in efficiency valuations. Eur J Oper Res. 2001;132:400–10.
Seiford LM, Zhu J. Modeling undesirable factors in efficiency evaluation. Eur J Oper Res. 2002;142(1):16–20.
Shephard RW. Theory of cost and production functions. Princeton, NJ: Princeton University Press; 1970.
Takamura T, Tone K. A comparative site evaluation study for relocating Japanese Government Agencies out of Tokyo. Soc Econ Plann Sci. 2003;37:85–102.
Thompson RG, Jr FD, Singleton RMThrall, Smith BA. Comparative site evaluation for locating a high-energy physics lab in Texas. Interfaces. 1986;16:35–49.
Thompson RG, Langemeier L, Lee C, Lee E, Thrall R. The role of multiplier bounds in efficiency analysis with application to Kansas farming. J Econ. 1990;46:93–108.
Tone K Sahoo BK. A reexamination of cost efficiency and cost elasticity in DEA. 2003. Working paper. National Graduate Institute for Policy Studies, Tokyo, Japan.
Wong Y-HB, Beasley JE. Restricting weight flexibility in data envelopment analysis. J Oper Res Soc. 1990;41:829–35.
Zhu J. DEA/AR analysis of the 1988–1989 performance of the Nanjing Textiles Corporation. Ann Oper Res. 1996a;66:311–35.
Zhu J. Data envelopment analysis with preference structure. J Oper Res Soc. 1996b;47(1):136–50.

Zhu J. Quantitative models for performance evaluation and benchmarking: data envelopment analysis with spreadsheets and DEA excel solver. Boston: Kluwer; 2003a.
Zhu J. Imprecise data envelopment analysis (IDEA): a review and improvement with an application. Eur J Oper Res. 2003b;144(3):513–29.
Zhu J. Quantitative models for performance evaluation and benchmarking: data envelopment analysis with spreadsheets. 2nd ed. Boston: Springer Science; 2009.

Chapter 2
Returns to Scale in DEA*

Rajiv D. Banker, William W. Cooper, Lawrence M. Seiford, and Joe Zhu

Abstract This chapter discusses returns to scale (RTS) in data envelopment analysis (DEA). The BCC and CCR models described in Chap. 1 of this handbook are treated in input-oriented forms, while the multiplicative model is treated in output-oriented form. (This distinction is not pertinent for the additive model, which simultaneously maximizes outputs and minimizes inputs in the sense of a vector optimization.) Quantitative estimates in the form of scale elasticities are treated in the context of multiplicative models, but the bulk of the discussion is confined to qualitative characterizations such as whether RTS is identified as increasing, decreasing, or constant. This is discussed for each type of model, and relations between the results for the different models are established. The opening section describes and delimits approaches to be examined. The concluding section outlines further opportunities for research and an Appendix discusses other approaches in DEA treatment of RTS.

Keywords Data envelopment analysis • Efficiency • Returns to scale

2.1 Introduction

It has long been recognized that Data Envelopment Analysis (DEA) by its use of mathematical programming is particularly adept at estimating inefficiencies in multiple input and multiple output production correspondences. Following Charnes, Cooper, and Rhodes (CCR 1978), a number of different DEA models

*Part of the material in this chapter is adapted from European Journal of Operational Research, Vol 154, Banker, R.D., Cooper, W.W., Seiford, L.M., Thrall, R.M. and Zhu, J, Returns to scale in different DEA models, 345–362, 2004, with permission from Elsevier Science.

J. Zhu (✉)
School of Business, Worcester Polytechnic Institute, Worcester, MA 01609, USA
e-mail: jzhu@wpi.edu

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_2,

have now appeared in the literature (see Cooper et al. 2000). During this period of model development, the economic concept of returns to scale (RTS) has also been widely studied within the different frameworks provided by these methods, and this is the topic to which this chapter is devoted.

In the literature of classical economics, RTS have typically been defined only for single-output situations. RTS are considered to be increasing if a proportional increase in all the inputs results in a more than proportional increase in the single output. Let α represent the proportional input increase and β represent the resulting proportional increase of the single output. Increasing returns to scale(IRS) prevail if $\beta > \alpha$, and decreasing returns to scale (DRS) prevail if $\beta < \alpha$. Banker (1984), Banker et al. (1984), and Banker and Thrall (1992) extend the RTS concept from the single-output case to multiple-output cases using DEA.

Two paths may be followed in treating RTS in DEA. The first path, developed by Färe, Grosskopf, and Lovell (FGL 1985, 1994), determines RTS by a use of ratios of radial measures. These ratios are developed from model pairs which differ only in whether conditions of convexity and subconvexity are satisfied. The second path stems from work by Banker (1984), Banker et al. (1984) and Banker and Thrall (1992). This path, which is the one we follow, includes, but is not restricted to, radial measure models. It extends to additive and multiplicative models as well, and does so in ways that provide opportunities for added insight into the nature of RTS and its treatment by the methods and concepts of DEA.

The FGL approach has now achieved a considerable degree of uniformity that has long been available – as in FGL (1985), for instance. See also FGL (1994). We therefore treat their approach in the Appendix to this chapter. This allows us to center this chapter on treating RTS with different models. These treatments have long been available but only in widely scattered literatures. We also delineate relations that have been established between these different treatments and extend this to relations that have also been established with the FGL approach. See Banker et al. (1996b), Zhu and Shen (1995), and Färe and Grosskopf (1994). In particular, Seiford and Zhu (1999) established the relations among these alternative approaches and provided a simple approach to RTS estimation without the need for checking multiple optimal solutions.

The plan of development in this chapter starts with a recapitulation of results from the very important paper by Banker and Thrall (1992). Although developed in the context of radial measure models, we also use the Banker and Thrall (1992) results to unify the treatment of all of the models we cover. This is done after we first cover the radial measure models that are treated by Banker and Thrall (1992). Proofs of their theorems are not supplied because these are already available in Banker and Thrall (1992). Instead refinements from Banker et al. (1996a) and from Banker et al. (1996b) are introduced, which are directed to (a) providing simpler forms for implementing the Banker–Thrall theorems and (b) eliminating some of the assumptions underlying these theorems.

We then turn to concepts such as the MPSS (Most Productive Scale Size) introduced by Banker (1984) to treat multiple output–multiple input cases

in DEA and extend RTS concepts built around the single-output case in classical economics. Additive and multiplicative models are then examined, and the latter are used to introduce (and prove) new theorems for determining scale elasticities.

The former (i.e., the additive case) is joined with a "goal vector" approach introduced by Thrall (1996a) to make contact with "invariance" and "balance" ideas that play prominent roles in the "dimensional analysis" used to guide the measurements used in the natural sciences (such as physics). We next turn to the class of multiplicative models where, as shown by Charnes et al. (1982, 1983) and Banker and Maindiratta (1986), the piecewise linear frontiers usually employed in DEA are replaced by a frontier that is piecewise Cobb–Douglas (= log linear). Scale elasticity estimates are then obtained from the exponents of these "Cobb–Douglas like" functions for the different segments that form a frontier, which need not be concave. A concluding section points up issues for further research.

The Appendix of this chapter presents the FGL approach. We then present a simple RTS approach developed by Zhu and Shen (1995) and Seiford and Zhu (1999) to avoid the need for checking the multiple optimal solutions. This approach will substantially reduce the computational burden because it relies on the standard CCR and BCC computational codes.

2.2 RTS Approaches with BCC Models

For ease of reference, we present here the BCC models. Suppose that we have n DMUs (decision-making units) where every DMU_j, $j = 1, 2, \ldots, n$, produces the same s outputs in (possibly) different amounts, y_{rj} ($r = 1, 2, \ldots, s$), using the same m inputs, x_{ij} ($i = 1, 2, \ldots, m$), also in (possibly) different amounts. The efficiency of a specific DMU_o can be evaluated by the "BCC model" of DEA in "envelopment form" as follows,

$$\begin{aligned}
&\min\ \theta_o - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right),\\
&\text{subject to}\\
&\theta_o x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- \qquad i = 1, 2, \ldots, m,\\
&y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ \qquad r = 1, 2, \ldots, s,\\
&1 = \sum_{j=1}^{n} \lambda_j,\\
&0 \le \lambda_j, s_i^-, s_r^+ \quad \forall\, i, r, j,
\end{aligned} \tag{2.1}$$

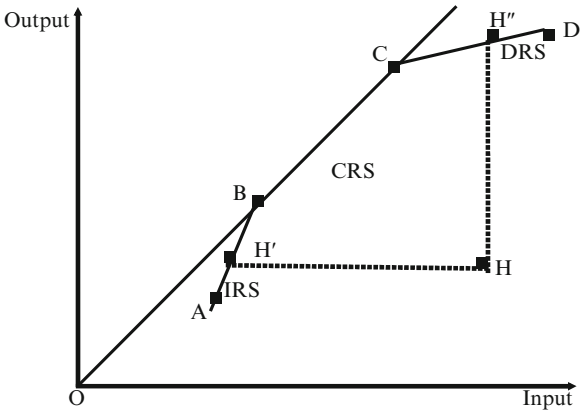


Fig. 2.1 Returns to scale

where, as discussed for the expression (1.6) in Chap. 1, $\varepsilon > 0$ is a non-Archimedean element defined to be smaller than any positive real number.

As noted in the abstract, we confine attention to input-oriented versions of these radial measure models and delay discussion of changes in mix, as distinct from changes in scale, until we come to the class of additive models where input and output orientations are treated simultaneously. Finally, we do use output orientations in the case of multiplicative models because, as will be seen, the formulations in that case do not create problems in distinguishing between scale and mix changes.

Remark. We should, however, note that input- and output-oriented models may give different results in their RTS findings. See Fig. 2.1 and related discussion below. Thus the result secured may depend on the orientation used. IRS may result from an input-oriented model, for example, while an application of an output-oriented model may produce a DRS characterization from the same data. See Golany and Yu (1994) for treatments of this problem.

The dual (multiplier) form of the BCC model represented in (2.1) is obtained from the same data that are then used in the following form:

$$\begin{aligned}
&\max z = \sum_{r=1}^{s} u_r y_{ro} - u_o, \\
&\text{subject to} \\
&\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} - u_o \le 0, \quad j = 1, \ldots, n, \\
&\qquad\qquad \sum_{i=1}^{m} v_i x_{io} \qquad = 1, \\
&v_i \ge \varepsilon,\ u_r \ge \varepsilon,\ u_o \text{ free in sign.}
\end{aligned} \tag{2.2}$$

The above formulations assume that $x_{ij}, y_{rj} \geq 0 \quad \forall\, i, r, j$. All variables in (2.2) are also constrained to be nonnegative – except for u_o, which may be positive, negative, or zero with consequences that make it possible to use optimal values of this variable to identify RTS.

When a DMU_o is efficient in accordance with the Definition 1.3 in Chap. 1, the optimal value of u_o, i.e., u_o^*, in (2.2), can be used to characterize the situation for RTS.

RTS generally has an unambiguous meaning only if DMU_o is on the efficiency frontier – since it is only in this state that a tradeoff between inputs and outputs is *required* to improve one or the other of these elements. However, there is no need to be concerned about the efficiency status in our analyses because efficiency can always be achieved as follows. If a DMU_o is not BCC efficient, we can use optimal values from (2.1) to project this DMU onto the BCC efficiency frontier via the following formulas,

$$\begin{cases} \hat{x}_{io} = \theta_o^* x_{io} - s_i^{-*} = \sum_{j=1}^{n} x_{ij}\lambda_j^*, & i = 1, \ldots, m, \\ \hat{y}_{ro} = y_{ro} + s_r^{+*} = \sum_{j=1}^{n} y_{rj}\lambda_j^*, & r = 1, \ldots, s, \end{cases} \tag{2.3}$$

where the symbol "*" denotes an optimal value. These are sometimes referred to as the "CCR Projection Formulas" because Charnes et al. (1978) showed that the resulting $\hat{x}_{io} \leq x_{io}$ and $\hat{y}_{ro} \geq y_{ro}$ correspond to the coordinates of a point on the efficiency frontier. They are, in fact, coordinates of the point used to evaluate DMU_o when (2.1) is employed.

Suppose that we have five DMUs, *A*, *B*, *C*, *D*, and H as shown in Fig. 2.1 from Zhu (2009). Ray OBC is the constant returns to scale (CRS) frontier. AB, BC, and CD constitute the BCC frontier, and exhibit increasing, constant, and decreasing returns to scale, respectively. *B* and *C* exhibit CRS. On the line segment AB, IRS prevail to the left of *B* for the BCC model and on the line segment CD, DRS prevail to the right of *C*. By applying (2.3) to point H, we have a frontier point H′ on the line segment AB where IRS prevail. However, if we use the output-oriented BCC model, the projection is onto H″ where DRS prevail. This is due to the fact that the input-oriented and the output-oriented BCC models yield different projection points on the BCC frontier and it is on the frontier that RTS is determined. See Zhu (2009) for discussion on "returns to scale regions."

We now present our theorem for RTS as obtained from Banker and Thrall (1992, p. 79) who identify RTS with the sign of u_o^* in (2.2) as follows:

Theorem 2.1. The following conditions identify the situation for RTS for the BCC model given in (2.2),

(i) IRS prevail at $(\hat{x}_o, \hat{y}_o)$ if and only if $u_o^* < 0$ for all optimal solutions.
(ii) DRS prevail at $(\hat{x}_o, \hat{y}_o)$ if and only if $u_o^* > 0$ for all optimal solutions.
(iii) Constant RTS prevail at $(\hat{x}_o, \hat{y}_o)$ if and only if $u_o^* = 0$ for at least one optimal solution.

Here, it may be noted, $(\hat{x}_o, \hat{y}_o)$ are the coordinates of the point on the efficiency frontier which is obtained from (2.3) in the evaluation of DMU_o via the solution to (2.1). Note, therefore, that a use of the projection makes it unnecessary to *assume* that the points to be analyzed are all on the BCC efficient frontier – as was assumed in Banker and Thrall (1992).

An examination of all optimal solutions can be onerous. Therefore, Banker and Thrall (1992) provide one way of avoiding a need for examining *all* optimal solutions. However, Banker et al. (1996a) approach is used here because it avoids the possibility of infinite solutions that are present in the Banker–Thrall approach. In addition, the Banker et al. (1996a) approach insures that the RTS analyses are conducted on the efficiency frontier. This is accomplished as follows.

Suppose that an optimum has been achieved with $u_o^* < 0$. As suggested by Banker et al. (1996a), the following model may then be employed to avoid having to explore all alternate optima,

$$
\begin{aligned}
&\text{maximize } \hat{u}_o,\\
&\text{subject to}\\
&\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} - \hat{u}_o \le 0, \quad j = 1, \ldots, n;\ j \neq 0,\\
&\sum_{r=1}^{s} u_r \hat{y}_{ro} - \sum_{i=1}^{m} v_i \hat{x}_{io} - \hat{u}_o \le 0, \quad j = 0,\\
&\qquad\qquad \sum_{i=1}^{m} v_i \hat{x}_{io} \quad = 1,\\
&\sum_{r=1}^{s} u_r \hat{y}_{ro} \qquad\qquad - \hat{u}_o = 1,\\
&\qquad v_i, u_r \ge 0 \quad \text{and} \quad \hat{u}_o \le 0,
\end{aligned}
\tag{2.4}
$$

where the $\hat{x}_{io}$ and $\hat{y}_{ro}$ are obtained from (2.3).

With these changes of data, the constraints for (2.4) are in the same form as (2.2) except for the added conditions $\sum_{r=1}^{s} u_r \hat{y}_{ro} - \hat{u}_o = 1$ and $\hat{u}_o \le 0$. The first of these conditions helps to ensure that we will be confined to the efficiency frontier. The second condition allows us to determine whether an optimal value can be achieved with $\max \hat{u}_o = 0$. If $\hat{u}_o^* = 0$ can be obtained, then condition (iii) of Theorem 2.1 is satisfied and RTS are constant. If, however, $\max \hat{u}_o = \hat{u}_o^* < 0$, then, as set forth in (i) of Theorem 2.1, RTS are increasing. In either case, the problem is resolved, and the need for examining all alternate optima is avoided in this way of implementing Theorem 2.1.

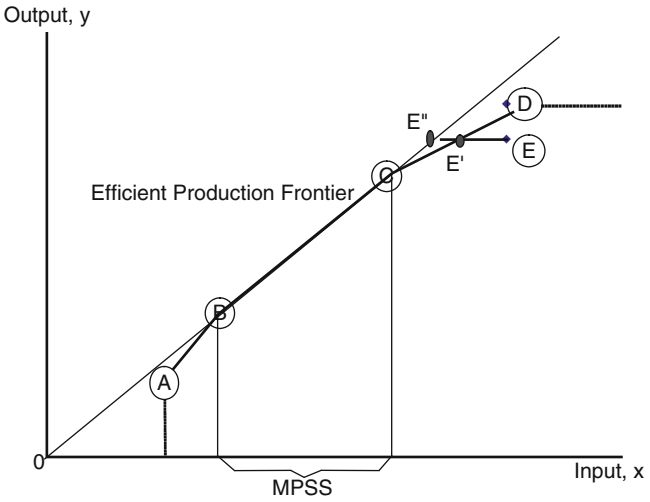


Fig. 2.2 Most productive scale size

We can deal in a similar manner with the case when $u_o^*>0$ by (a) reorienting the objective in (2.4) to "minimize" $\hat{u}_o$ and (b) replacing the constraint $\hat{u}_o \leq 0$ with $\hat{u}_o \geq 0$. All other elements of (2.4) remain the same and if min $\hat{u}_o = \hat{u}_o^* > 0$ then condition (ii) of Theorem 2.1 is applicable while if min $\hat{u}_o = \hat{u}_o^* = 0$ then condition (iii) is applicable.

Reference to Fig. 2.2 can help us to interpret these results. This figure portrays the case of one input, x, and one output, y. The coordinates of each point are listed in the order (x, y). Now, consider the data for A that has the coordinates $(x = 1, y = 1)$, as shown at the bottom of Fig. 2.2. The "supports" at A form a family that starts at the vertical line (indicated by the dotted line) and continues through rotations about A until coincidence is achieved with the line connecting A to B. All of these supports will have negative intercepts so $u_o^* < 0$ and the situation for A is one of IRS as stated in (i) of Theorem 2.1.

The reverse situation applies at D. Starting with the horizontal line indicated by the dots, supports can be rotated around D until coincidence is achieved with the line connecting D and C. In all cases, the intercept is positive so $u_o^* > 0$ and RTS are decreasing as stated in (ii) of Theorem 2.1.

Rotations at C or B involve a family of supports in which at least one member will achieve coincidence with the broken line going through the origin so that, in at least this single case, we will have $u_o^* = 0$, in conformance with the condition for constant RTS in (iii) of Theorem 2.1.

Finally, we turn to E, the only point that is BCC inefficient in Fig. 2.2. Application of (2.3), however, projects E into E' – a point on the line between C and D – and, therefore, gives the case of DRS with a unique solution of $\hat{u}_o^* > 0$.

Hence, all possibilities are comprehended by Theorem 2.1 for the qualitative RTS characterizations that are of concern here. Thus, for these characterizations only the signs of the nonzero values of $\hat{u}_o^*$ suffice.

2.3 RTS Approaches with CCR Models

We now turn to the CCR models that, as discussed in Chap. 1 of this handbook, take the following form,

$$\begin{aligned}
&\text{minimize } \theta - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right),\\
&\text{subject to}\\
&\theta x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-,\\
&y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+,\\
&0 \le \lambda_j, s_i^-, s_r^+ \quad \forall\, i, j, r. \qquad (2.5)
\end{aligned}$$

As can be seen, this model is the same as the "envelopment form" of the BCC model in (2.1) except for the fact that the condition $\sum_{j=1}^{n} \lambda_j = 1$ is omitted. In consequence, the variable u_o, which appears in the "multiplier form" for the BCC model in (2.2), is omitted from the dual (multiplier) form of this CCR model. The projection formulas expressed in (2.3) are the same for both models. We can, therefore, use these same projections to move all points onto the efficient frontier for (2.5) and proceed directly to RTS characterizations for (2.5), which are supplied by the following theorem from Banker and Thrall (1992).

Theorem 2.2. The following conditions identify the situation for RTS for the CCR model given in (2.5):

(i) CRS prevail at $(\hat{x}_o, \hat{y}_o)$ if $\sum \lambda_j^* = 1$ in any alternate optimum.
(ii) DRS prevail at $(\hat{x}_o, \hat{y}_o)$ if $\sum \lambda_j^* > 1$ for all alternate optima.
(iii) IRS prevail at $(\hat{x}_o, \hat{y}_o)$ if $\sum \lambda_j^* < 1$ for all alternate optima.

Following Banker et al. (1996b), we can avoid the need for examining all alternate optima. This is done as follows. Suppose that an optimum has been obtained for (2.5) with $\sum \lambda_j^* < 1$. We then replace (2.5) with

$$\text{maximize} \sum_{j=1}^{n} \hat{\lambda}_j + \varepsilon \left(\sum_{i=1}^{m} \hat{s}_i^- + \sum_{r=1}^{s} \hat{s}_r^+ \right),$$

subject to

$$\theta^* x_{io} = \sum_{j=1}^{n} x_{ij} \hat{\lambda}_j + \hat{s}_i^- \quad \text{for } i = 1, \ldots, m,$$

$$y_{ro} = \sum_{j=1}^{n} y_{ri} \hat{\lambda}_j - \hat{s}_r^+ \quad \text{for } r = 1, \ldots, s,$$

$$1 \geq \sum_{j=1}^{n} \hat{\lambda}_j,$$

$$\text{with } 0 \leq \hat{\lambda}_j, \hat{s}_i^-, \hat{s}_r^+ \quad \forall\, i, j, r, \tag{2.6}$$

where θ^* is the optimal value of θ secured from (2.5).

Remark. This model can also be used for setting scale-efficient targets when multiple optimal solutions in model (2.5) are present. See Zhu (2000, 2002).

We note for (2.5) that we may omit the two-stage process described for the CCR model in Chap. 1 – i.e., the process in which the sum of the slacks are maximized in stage 2 after θ^* has been determined. This is replaced with a similar two-stage process for (2.6) because only the optimal value of θ is needed from (2.5) to implement the analysis now being described. The optimal solution to (2.6) then yields values of $\hat{\lambda}_j^*, j = 1, \ldots, n$, for which the following theorem is immediate,

Theorem 2.3. Given the existence of an optimal solution with $\sum \lambda_j^* < 1$ in (2.5), the RTS at $(\hat{x}_o, \hat{y}_o)$ are constant if and only if $\sum \hat{\lambda}_j^* = 1$ and RTS are increasing if and only if $\sum \hat{\lambda}_j^* < 1$ in (2.6).

Consider $A = (1,\ 1)$ as shown at the bottom of Fig. 2.2. Because we are only interested in θ^*, we apply (1.4) in Chap. 1 to obtain

$$\text{minimize } \theta,$$

subject to

$$1\theta \geq 1\lambda_A + \frac{3}{2}\lambda_B + 3\lambda_C + 4\lambda_D + 4\lambda_E,$$

$$1 \leq 1\lambda_A + 2\lambda_B + 4\lambda_C + 5\lambda_D + \frac{9}{2}\lambda_E,$$

$$0 \leq \lambda_A, \lambda_B, \lambda_C, \lambda_D, \lambda_E. \tag{2.7}$$

This problem has $\theta^* = 3/4$ and hence A is found to be inefficient. Next, we observe that this problem has alternate optima because this same $\theta^* = 3/4$ can be obtained from either $\lambda_B^* = 1/2$ or from $\lambda_C^* = 1/4$ with all other $\lambda^* = 0$. For each of these optima, we have $\sum \lambda_j^* < 1$, so we utilize (2.6) and write

$$
\begin{aligned}
&\text{maximize } \hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E + \varepsilon(\hat{s}^- + s^+),\\
&\text{subject to}\\
&\frac{3}{4} = 1\hat{\lambda}_A + \frac{3}{2}\hat{\lambda}_B + 3\hat{\lambda}_C + 4\hat{\lambda}_D + 4\hat{\lambda}_E + \hat{s}^-,\\
&1 = 1\hat{\lambda}_A + 2\hat{\lambda}_B + 4\hat{\lambda}_C + 5\hat{\lambda}_D + \frac{9}{2}\hat{\lambda}_E - \hat{s}^+,\\
&1 \geq \hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E,\\
&0 \leq \hat{\lambda}_A, \hat{\lambda}_B, \hat{\lambda}_C, \hat{\lambda}_D, \hat{\lambda}_E.
\end{aligned} \tag{2.8}
$$

so that $\sum \hat{\lambda}_j^* \equiv \hat{\lambda}_A^* + \hat{\lambda}_B^* + \hat{\lambda}_C^* + \hat{\lambda}_D^* + \hat{\lambda}_E^*$ with all $\hat{\lambda}$ nonnegative. An optimal solution is $\hat{\lambda}_B^* = 1/2$ and all other $\hat{\lambda}^* = 0$. Hence, $\sum \hat{\lambda}_j^* < 1$, so from Theorem 2.3, IRS prevail at A.

We are here restricting attention to solutions of (2.6) with $\sum_{j=1}^{n} \hat{\lambda}_j \leq 1$, as in the constraint of (2.6), but the examples we provide below show how to treat situations in which θ^* is associated with solutions of (2.5) that have values $\sum \lambda_j^* > 1$.

Consider $E = (4, 9/2)$ for (2.6), as a point that is not on either (i) the BCC efficiency frontier represented by the solid lines in Fig. 2.2 or (ii) the CCR efficiency frontier represented by the broken line from the origin. Hence, both the BCC and CCR models find E to be inefficient. Proceeding via the CCR envelopment model in (2.5) with the slacks omitted from the objective, we get

$$
\begin{aligned}
&\text{minimize } \theta,\\
&\text{subject to}\\
&4\theta \geq 1\lambda_A + \frac{3}{2}\lambda_B + 3\lambda_C + 4\lambda_D + 4\lambda_E,\\
&\frac{9}{2} \leq 1\lambda_A + 2\lambda_B + 4\lambda_C + 5\lambda_D + \frac{9}{2}\lambda_E,\\
&0 \leq \lambda_A, \lambda_B, \lambda_C, \lambda_D, \lambda_E.
\end{aligned} \tag{2.9}
$$

Again we have alternate optima with, now, $\theta^* = 27/32$ for either $\lambda_B^* = 9/4$ or $\lambda_C^* = 9/8$ and all other $\lambda^* = 0$. Hence, in both cases, we have $\sum \hat{\lambda}_j^* > 1$. Continuing in an obvious way, we next reorient the last constraint and the objective in (2.6) to obtain

$$
\begin{aligned}
&\text{minimize } \left(\hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E\right) - \varepsilon(\hat{s}^- + s^+),\\
&\text{subject to}\\
&\frac{27}{8} = 1\hat{\lambda}_A + \frac{3}{2}\hat{\lambda}_B + 3\hat{\lambda}_C + 4\hat{\lambda}_D + 4\hat{\lambda}_E + \hat{s}^-,\\
&\frac{9}{2} = 1\hat{\lambda}_A + 2\hat{\lambda}_B + 4\hat{\lambda}_C + 5\hat{\lambda}_D + \frac{9}{2}\hat{\lambda}_E - \hat{s}^+,\\
&1 \leq \hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E,\\
&0 \leq \hat{\lambda}_A, \hat{\lambda}_B, \hat{\lambda}_C, \hat{\lambda}_D, \hat{\lambda}_E.
\end{aligned} \tag{2.10}
$$

This has its optimum at $\hat{\lambda}_C^* = 9/8$ with all other $\hat{\lambda}^* = 0$. So, in conformance with Theorem 2.3, as given for (2.6), we associate E with DRS.

There is confusion in the literature on the RTS characterizations obtained from Theorems 2.1 and 2.2 and the BCC and the CCR models with which they are associated. Hence, we proceed further as follows.

As noted earlier, RTS generally has an unambiguous meaning only for points on the efficiency frontier. When the BCC model as given in (2.1) is used on the data in Fig. 2.2, the primal model projects E into E' with coordinates (7/2, 9/2) on the segment of the line $y = 1 + x$, which connects C to D on the BCC efficiency frontier. Comparing this with $E = (4, 9/2)$ identifies E as having an inefficiency in the amount of 1/2 unit in its input. This is a technical inefficiency, in the terminology of DEA. Turning to the dual for E formed from the BCC model, as given in (2.2), we obtain $u_o^* = 1/4$. Via Theorem 2.1 this positive value of u_o^* suggests that RTS are either decreasing or constant at $E' = (28/8, 9/2)$ – the point to which E is projected to obtain access to model (2.4). Substitution in the latter model yields a value of $\hat{u}_o^* \doteq 2/7$, which is also positive, thereby identifying E' with the DRS that prevail for the BCC model on this portion of the efficiency frontier in Fig. 2.2.

Next, we turn to the conditions specified in Theorem 2.2, which are identified with the CCR envelopment model (2.5). Here, we find that the projection is to a new point $E'' = (27/8, 9/2)$, which is on the line $y = 4/3x$ corresponding to the broken line from the origin that coincides with the segment from B to C in Fig. 2.2. This ray from the origin constitutes the efficiency frontier for the CCR model, which, when used in the manner we have previously indicated, simultaneously evaluates the technical and RTS performances of E. In fact, as can be seen from the solution to (2.9), this evaluation is effected by either $\hat{\lambda}_B^* = 9/4$ or $\hat{\lambda}_C^* = 9/8$ – which are variables associated with vectors in a "CRS region" that we will shortly associate with "most productive scale size" (MPSS) for the BCC model. The additional 1/8 unit input reduction effected in going from E'' to E'' is needed to adjust to the efficient mix that prevails in this MPSS region which the CCR model is using to evaluate E.

Thus, the CCR model as given in (2.5) simultaneously evaluates scale and purely technical inefficiencies, while the BCC model, as given in (2.1), separates out the scale inefficiencies for evaluation in its associated dual (=multiplier) form as given in (2.2). Finally, as is well known, a simplex method solution to (2.1) automatically supplies the solution to its dual in (2.2). Thus, no additional computations are required to separate the purely technical inefficiency characterizations obtained form (2.1) and the RTS characterizations obtained from (2.2). Both sets of values are obtainable from a solution to (2.1).

We now introduce the following theorem that allows us to consider the relations between Theorems 2.1 and 2.2 in the RTS characterization.

Theorem 2.4. Suppose that DMU_o is designated as efficient by the CCR model, DMU_o then it is also designated as efficient by the BCC model.

Proof. The CCR and BCC models differ only because the latter has the additional constraint $\sum_{j=1}^{n} \lambda_j = 1$. The following relation must therefore hold

$$\theta^*_{\text{CCR}} - \varepsilon\left(\sum_{i=1}^{m} s_i^{-*} + \sum_{r=1}^{s} s_r^{+*}\right) \leq \theta^*_{\text{BCC}} - \varepsilon\left(\sum_{i=1}^{m} s_i^{-*} + \sum_{r=1}^{s} s_r^{+*}\right),$$

where the expressions on the left and right of the inequality respectively designate optimal values for objective of the CCR and BCC models.

Now, suppose DMU_o is found to be efficient with the CCR model. This implies $\theta^*_{\text{CCR}} = 1$ and all slacks are zero for the expression on the left. Hence, we will have

$$1 \leq \theta^*_{\text{BCC}} - \varepsilon\left(\sum_{i=1}^{m} s_i^{-*} + \sum_{r=1}^{s} s_r^{+*}\right).$$

However, the x_{io} and y_{ro} values appear on both the left and right sides of the corresponding constraints in the DEA models. Hence, choosing $\lambda^*_o = \lambda^*_j = 1$, we can always achieve equality with θ^*_{BCC} which is the lower bound in this exact expression with all slacks zero. Thus, this DMU_o will also be characterized as efficient by the BCC model whenever it is designated as efficient by the CCR model. ■

We now note that the reverse of this theorem is not true. That is, a DMU_o may be designated as efficient by the BCC model but not by the CCR model. Even when both models designate a DMU_o as inefficient, moreover, the measures of inefficiency may differ. Application of the BCC model to point E in Fig. 2.2, for example, will designate E'' on the line connecting C and D to evaluate its efficiency. However, utilization of the CCR model will designate E' with $\theta^*_{\text{CCR}} < \theta^*_{\text{BCC}}$ so that $1 - \theta^*_{\text{CCR}} > 1 - \theta^*_{\text{BCC}}$, which shows a greater inefficiency value for DMU_o when the CCR model is used.

Because DEA evaluates relative efficiency, it will always be the case that at least one DMU will be characterized as efficient by either model. However, there will always be at least one point of intersection between these two frontiers. Moreover, the region of the intersection will generally expand a DMU set to be efficient with the CCR model. The greatest spread between the envelopments will then constitute extreme points that define the boundaries of the intersection between the CCR and BCC models.

The way Theorem 2.2 effects its efficiency characterization is by models of linear programming algorithms that use "extreme point" methods. That is, the solutions are expressed in terms of "basis sets" consisting of extreme points. The extreme points B and C in Fig. 2.2 can constitute active members of such an "optimal basis" where, by "active member," we refer to members of a basis that have nonzero coefficients in an optimal solution.

Now, as shown in Cooper et al. (2000), active members of an optimal basis are necessarily efficient. For instance, $\lambda^*_B = 1/2$ in the solution to (2.8) designates B as

an active member of an optimal basis and the same is true for $\lambda_C^* = 1/4$, and both B and C are therefore efficient. In both cases, we have $\sum \lambda_j^* < 1$ and we have IRS at point (3/4, 1) on the CRS ray, which is used to evaluate A. In other words, $\sum \lambda_j^* < 1$ shows that B and C both lie below the region of intersection because its coordinates are smaller in value than the corresponding values of the active members in the optimal basis.

Turning to the evaluation of E, we have $\theta^* = 27/32 < 1$ showing that E is inefficient in (2.10). Either $\lambda_B^* = 9/4$ or $\lambda_C^* = 9/8$ can serve as active member in the basis. Thus, to express E'' in terms of these bases, we have

$$\underset{E''}{\left(\frac{27}{8},\frac{9}{2}\right)} = \underset{B}{\frac{9}{4}\left(\frac{3}{2},2\right)} = \underset{C}{\frac{9}{8}(3,4)}$$

because E'' lies above the region of intersection, as shown by $\sum \lambda_j^* > 1$ for either of the optimal solutions.

As shown in the next section of this chapter, Banker (1984) refers to the region between B and C as the region of most productive scale size (MPSS). For justification, we might note that the slope of the ray from the origin through B and C is steeper than the slope of any other ray from the origin that intersects the production possibility set (PPS). This measure with the output per unit input is maximal relative to any other ray that intersects PPS.

Hence, Theorem 2.2 is using the values of $\sum \lambda_j^*$ to determine whether RTS efficiency has achieved MPSS and what needs to be done to express this relative to the region of MPSS. We can, therefore, conclude with a corollary to Theorem 2.4: DMU_o will be at MPSS if and only if $\sum \lambda_j^* = 1$ in an optimal solution when it is evaluated by a CCR model.

To see how this all comes about mathematically and how it relates to the RTS characterization, we note that the optimal solution for the CCR model consists of all points on the ray from the origin that intersect the MPSS region. If the point being evaluated is in MPSS, it can be expressed as a convex combination of the extreme points of MPSS so that $\sum \lambda_j^* = 1$. If the point is above the region, its coordinate values will all be larger than their corresponding coordinates in MPSS so that we will have $\sum \lambda_j^* > 1$. If the point is below the region, we will have $\sum \lambda_j^* < 1$. Because the efficient frontier, as defined by the BCC model, is strictly concave, the solution will designate this point as being in the region of constant, decreasing, or increasing RTS, respectively.

Thus, the CCR model simultaneously evaluates RTS and technical inefficiency while the BCC model separately evaluates technical efficiency with θ_{BCC}^* from the envelopment model and RTS with u_o^* obtained from the multiplier model. As Fig. 2.2 illustrates, at point E, the evaluation for the CCR model is global with RTS always evaluated relative to MPSS. The evaluation for the BCC model is local with u_o^* being determined by the facet of the efficient frontier in which the

point used to evaluate DMU_o is located. As a consequence, it always be the case that $\theta^*_{CCR} < \theta^*_{BCC}$ unless the point used to evaluate DMU_o is in the region of MPSS, in which case $\theta^*_{CCR} = \theta^*_{BCC}$ will obtain.

2.4 Most Productive Scale Size

There is some ambiguity in dealing with points like *B* and *C* in Fig. 2.2 because the condition that prevails depends on the direction in which movement is to be effected. As noted by Førsund (1996) this situation was dealt with by Ragnar Frisch – who pioneered empirical studies of production and suggested that the orientation should be toward maximizing the output per unit input when dealing with technical conditions of efficiency. See Frisch (1964). However, Frisch (1964) dealt only with the case of single outputs. Extensions to multiple output–multiple input situations can be dealt with by the concept of Most Productive Scale Size (MPSS) as introduced into the DEA literature by Banker (1984). To see what this means consider

$$(X_o\alpha, Y_o\beta), \tag{2.11}$$

with $\beta, \alpha \geq 0$ representing scalars and X_o and Y_o representing input and output vectors, respectively, for DMU_o. We can continue to move toward a possibly better (i.e., more productive) RTS situation as long as $\beta/\alpha \neq 1$. In other words, we are not at a point which is MPSS when either (a) all outputs can be increased in proportions that are at least as great as the corresponding proportional increases in all inputs needed to bring them about, or (b) all inputs can be decreased in proportions that are at least as great as the accompanying proportional reduction in all outputs. Only when $\beta/\alpha = 1$, or $\alpha = \beta$, will RTS be constant, as occurs at MPSS.

One way to resolve problems involving returns to scale for multiple output–multiple input situations would use a recourse to prices, costs (or similar weights) to determine a "best" or "most economical" scale size. Here, however, we are using the concept of MPSS in a way that avoids the need for additional information on unit prices, costs, etc., by allowing all inputs and outputs to vary simultaneously in the proportions prescribed by α and β in (2.11). Hence, MPSS allows us to continue to confine attention to purely technical inefficiencies, as before, while allowing for other possible choices after scale changes and size possibilities have been identified and evaluated in our DEA analyses.

The interpretation we have just provided for (2.11) refers to RTS locally, as is customary – e.g., in economics. However, this does not exhaust the uses that can be made of Banker's (1984) MPSS. For instance, we can now replace our preceding local interpretation of (2.11) by one which is oriented globally. That is, we seek to characterize the returns to scale conditions for DMU_o with respect to MPSS instead of restricting this evaluation to the neighborhood of the point (X_o, Y_o) where, say, a derivative is to be evaluated. See Varian (1984, p. 20) for economic interpretations of restrictions needed to justify uses of derivatives. We also do

this in a way that enables us to relate Theorems 2.1 and 2.2 to each other and thereby provide further insight into how the BCC and CCR models relate to each other in scale size (and other) evaluations.

For these purposes, we introduce the following formulation,

$$\begin{aligned}
&\text{maximize } \frac{\beta}{\alpha},\\
&\text{subject to}\\
&\beta Y_o \le \sum_{j=1}^{n} Y_j \lambda_j,\\
&\alpha X_o \ge \sum_{j=1}^{n} X_j \lambda_j,\\
&1 = \sum_{j=1}^{n} \lambda_j,\\
&0 \le \beta, \alpha, \lambda_j, \quad j = 1, \ldots, n.
\end{aligned} \tag{2.12}$$

Now note that the condition $\sum \lambda_j = 1$ appears just as it does in (2.1). However, in contrast to (2.1), we are now moving to a global interpretation by jointly maximizing the proportional increase in outputs and minimizing the proportional decrease in inputs. We are also altering the characterizations so that these α and β values now yield new vectors $\hat{X}_o = \alpha X_o$ and $\hat{Y}_o = \beta Y_o$, which we can associate with points which are MPSS, as in the following

Theorem 2.5. A necessary condition for DMU_o, with output and input vectors Y_o and X_o, to be MPSS is max $\beta/\alpha = 1$ in (2.12), in which case RTS will be constant.

Theorem 2.5 follows from the fact that $\beta = \alpha = 1$ with $\lambda_j = 0$, $\lambda_o = 1$ for $j \neq o$ is a solution of (2.12), so that, always, max $\beta/\alpha = \beta^*/\alpha^* \ge 1$. See the appendix in Cooper et al. (1996) for a proof and a reduction of (2.12) to a linear programming equivalent.

We illustrate with $D = (4,5)$ in Fig. 2.2 for which we utilize (2.12) to obtain

$$\begin{aligned}
&\text{Maximize } \frac{\beta}{\alpha},\\
&\text{subject to}\\
&5\beta \le 1\lambda_A + 2\lambda_B + 4\lambda_C + 5\lambda_D + \frac{9}{2}\lambda_E,\\
&4\alpha \ge 1\lambda_A + \frac{3}{2}\lambda_B + 3\lambda_C + 4\lambda_D + 4\lambda_E,\\
&1 = \lambda_A + \lambda_B + \lambda_C + \lambda_D + \lambda_E,\\
&0 \le \lambda_A, \lambda_B, \lambda_C, \lambda_D, \lambda_E.
\end{aligned} \tag{2.13}$$

This has an optimum at $\lambda_B^* = 1$ with $\alpha^* = 3/8$ and $\beta^* = 2/5$ to give $\beta^*/\alpha^* = 16/15 > 1$. Thus, MPSS is not achieved. Substituting in (2.13) with $\lambda_B^* = 1$, we can use this solution to obtain $4\alpha^* = 3/2$ and $5\beta^* = 2$ which are the coordinates of B in Fig. 2.2. Thus, $D = (4,5)$ is evaluated globally by reference to $B = (3/2,2)$, which is in the region of CRS and hence is MPSS.

There is also an alternate optimum to (2.13) with $\lambda_C^* = 1$ and $\alpha^* = 3/4$, $\beta^* = 4/5$ so, again, $\beta^*/\alpha^* = 16/15$, and D is not at MPSS. Moreover, $4\alpha^* = 5$, $5\beta^* = 4$ gives the coordinates of $C = (3, 4)$. Thus, D is again evaluated globally by a point in the region of MPSS. Indeed, any point in this region of MPSS would give the same value of $\beta^*/\alpha^* = 16/15$, since all such points are representable as convex combinations of B and C.

Theorem 2.6. Sign conditions for BCC and CCR models:

(i) The case of IRS. $u_o^* < 0$ for all optimal solutions to (2.2) if and only if $\left(\sum \lambda_j^* - 1\right) < 0$ for all optimal solutions to (2.5).

(ii) The case of DRS. $u_o^* > 0$ for all optimal solutions to (2.2) if and only if $\left(\sum \lambda_j^* - 1\right) > 0$ for all optimal solutions to (2.5).

(iii) The case of CRS. $u_o^* = 0$ for some optimal solutions to (2.2) if and only if $\left(\sum \lambda_j^* - 1\right) = 0$ for some optimal solution to (2.5).

This theorem removes the possibility that uses of the CCR and BCC models might lead to different RTS characterizations. It is also remarkable because differences might be expected from the fact that (2.2) effects its evaluations locally with respect to a neighboring facet while (2.5) effects its evaluations globally with respect to a facet (or point) representing MPSS.

To see what this means we focus on active members of an optimal solution set as follows.

Turning to E in Fig. 2.2 we see that it is evaluated by E' when (2.1) is used. This point, in turn, can be represented as a convex combination of C and D with both of the latter vectors constituting active members of the optimal basis. The associated support coincides with the line segment connecting C and D with a (unique) value $u_o^* > 0$ so RTS are decreasing, as determined from (2.2). This is a local evaluation. When (2.5) is used, the projection is to E'', with alternate optima at B or C respectively serving as the only active member of the optimal basis. Hence the evaluation by the CCR model is effected globally. Nevertheless, the same DRS characterization is secured.

We now note that E'' may be projected into the MPSS region by means of the following formulas,

$$\frac{\theta^* x_{io} - s_i^{-*}}{\sum_{j=1}^{n} \hat{\lambda}_j^*},$$

$$\frac{y_{ro} + s_i^{+*}}{\sum_{j=1}^{n} \hat{\lambda}_j^*}, \tag{2.14}$$

where the denominators are secured from (2.6). This convexification of (2.3), which is due to Banker and Morey (1986), provides a different projection than (2.3). We illustrate for E'' by using the solutions for (2.9) to obtain

$$\frac{4\theta^* - s_i^{-*}}{9/4} = \frac{27/8}{9/4} = 3/2,$$

$$\frac{y_{ro} + s_i^{+*}}{9/4} = \frac{9/2}{9/4} = 2.$$

This gives the coordinates of B from one optimal solution. The other optimal solution yields the coordinates of C via

$$\frac{4\theta^* - s_i^{-*}}{9/8} = \frac{27/8}{9/8} = 3,$$

$$\frac{y_{ro} + s_i^{+*}}{9/8} = \frac{9/2}{9/8} = 4.$$

This additional step brings us into coincidence with the results already described for the MPSS model given in (2.13). Consistency is again achieved even though the two models proceed by different routes. The MPSS model in (2.12) bypasses the issue of increasing vs. decreasing RTS and focuses on the issue of MPSS, but this same result can be achieved for (2.5) by using the additional step provided by the projection formula (2.14).

2.5 Additive Models

The model (2.12), which we used for MPSS, avoids the problem of choosing between input and output orientations, but this is not the only type of model for which this is true. The additive models to be examined in this section also have this property. That is, these models simultaneously maximize outputs and minimize inputs, in the sense of vector optimizations.

The additive model we select is

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

subject to

$$\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = x_{io}, \quad i = 1, 2, \ldots, m,$$

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

$$\sum_{j=1}^{n} \lambda_j = 1,$$

$$\lambda_j, s_i^-, s_r^+ \geq 0. \tag{2.15}$$

This model utilizes the "goal vector" approach of Thrall (1996a) in which the slacks in the objective are accorded "goal weights" which may be subjective or objective in character. Here we want to use these "goal weights" to ensure that the units of measure associated with the slack variables do not affect the optimal solution choices.

Employing the language of "dimensional analysis," as in Thrall (1996a), we want these weights to be "contragredient" to insure that the resulting objective will be "dimensionless." That is, we want the solutions to be free of the dimensions in which the inputs and outputs are stated. An example is the use of the input and output ranges in Cooper et al. (1999) to obtain $g_i = 1/R_i^-$, $g_r = 1/R_r^+$ where R_i^- is the range for the ith input and R_r^+ is the range for the rth output. This gives each term in the objective of (2.15) a contragredient weight. The resulting value of the objective is dimensionless, as follows from the fact that the s_i^- and s_r^+ in the numerators are measured in the same units as the R_i^- and R_r^+ in the denominators. Hence the units of measure cancel.

The condition for efficiency given in Definition 1.3 in Chap. 1 for the CCR model is now replaced by the following simpler condition,

Definition 2.1. A DMU$_o$ evaluated by (2.15) is efficient if and only if all slacks are zero.

Thus, in the case of additive models it suffices to consider only condition (ii) in Definition 1.3. Moreover this condition emerges from the second stage solution procedure associated with the non-Archimedean $\varepsilon > 0$ in (1.1). Hence we might expect that RTS characterizations will be related, as we see now.

To start our RTS analyses for these additive models we first replace the CCR projections of (2.3) with

$$\hat{x}_{io} = x_{io} - s_i^{-*}, \quad i = 1, \ldots, m,$$
$$\hat{y}_{ro} = y_{ro} + s_r^{+*}, \quad r = 1, \ldots, s, \tag{2.16}$$

where s_i^{-*} and s_r^{+*} are optimal slacks obtained from (2.15). Then we turn to the dual (multiplier) model associated with (2.15) which we write as follows,

$$\begin{aligned}
&\min \sum_{i=1}^{m} v_i x_{io} - \sum_{r=1}^{s} \mu_{ri} y_{ro} + u_o, \\
&\text{subject to} \\
&\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_{ri} y_{rj} + u_o \geq 0, \quad j = 1, \ldots, m, \\
&v_i \geq g_i^-, \ \mu_r \geq g_r^+; \ u_o \text{ free.}
\end{aligned} \tag{2.17}$$

We are thus in position to use Theorem 2.1 for "additive" as well as "radial measures" as reflected in the BCC and CCR models discussed in earlier parts of this chapter. Hence we again have recourse to this theorem where, however, we note the difference in objectives between (2.2) and (2.17), including the change from $-u_o$ to $+u_o$. As a consequence of these differences we also modify (2.4) to the following,

$$\begin{aligned}
&\text{Maximize } \hat{u}_o, \\
&\text{subject to} \\
&\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} - \hat{u}_o \leq 0, \quad j = 1, \ldots, m; \ j \neq 0, \\
&\sum_{r=1}^{s} \mu_r \hat{y}_{ro} - \sum_{i=1}^{m} v_i \hat{x}_{io} - \hat{u}_o = 0, \\
&\mu_r \geq g_r^+, \ v_i \geq g_i^-, \ \hat{u}_o \leq 0.
\end{aligned} \tag{2.18}$$

Here we have assumed that $u_o^* < 0$ was achieved in a first-stage use of (2.17). Hence, if $\hat{u}_o^* < 0$ is maximal in (2.18) then RTS are increasing at $(\hat{x}_o, \hat{y}_o)$ in accordance with (i) in Theorem 2.1 whereas if $\hat{u}_o^* = 0$ then (iii) applies and RTS are constant at this point $(\hat{x}_o, \hat{y}_o)$ on the efficiency frontier.

For $u_o^*>0$ in stage one, the objective and the constraint on $\hat{u}_o$ are simply reoriented in the manner we now illustrate by using (2.15) to evaluate E in Fig. 2.2 via

$$\begin{aligned}
&\max s^- + s^+, \\
&\text{subject to} \\
&\lambda_A + \frac{3}{2}\lambda_B + 3\lambda_C + 4\lambda_D + 4\lambda_E + s^- = 4, \\
&\lambda_A + 2\lambda_B + 4\lambda_C + 5\lambda_D + \frac{9}{2}\lambda_E - s^+ = \frac{9}{2}, \\
&\lambda_A + \lambda_B + \lambda_C + \lambda_D + \lambda_E = 1, \\
&s^-, s^+, \lambda_A, \lambda_B, \lambda_C, \lambda_D, \lambda_E \geq 0,
\end{aligned}$$

where we have used unit weights for the g_i^-, g_r^+, to obtain the usual additive model formulation. (See Thrall (1996b) for a discussion of the applicable condition

for a choice of such "unity" weights.) This has an optimal solution with $\lambda_C^* = \lambda_D^* = s^{-*} = 1/2$ and all other variables zero. To check that this is optimal we turn to the corresponding dual (multiplier) form for the above envelopment model, which is

$$\begin{aligned}
&\min 4v - \frac{9}{2}\mu + u_o,\\
&\text{subject to}\\
&v - \ \mu \ + u_o \geq 0,\\
&\frac{3}{2}v - 2\mu \ + u_o \geq 0,\\
&3v - 4\mu \ + u_o \geq 0,\\
&4v - 5\mu \ + u_o \geq 0,\\
&4v - \frac{9}{2}\mu \ + u_o \geq 0,\\
&v, \mu \geq 1, \quad u_o \text{ free.}
\end{aligned}$$

The solution $v^* = \mu^* = u_o^* = 1$ satisfies all constraints and gives $4v^* - (9/2)\mu^* + u_o^* = 1/2$. This is the same value as in the preceding problem so that, by the dual theorem of linear programming, both solutions are optimal.

To determine the conditions for RTS, we use (2.16) to project E into E' with coordinates $(\hat{x}, \hat{y}) = (7/2,\ 9/2)$ in Fig. 2.2. Then we utilize the following reorientation of (2.18),

$$\begin{aligned}
&\min \hat{u}_o,\\
&\text{subject to}\\
&v - \ \mu \ + \hat{u}_o \geq 0,\\
&\frac{3}{2}v - 2\mu + \hat{u}_o \geq 0,\\
&3v - 4\mu \ + \hat{u}_o \geq 0,\\
&4v - 5\mu + \hat{u}_o \geq 0,\\
&\frac{7}{2}v - \frac{9}{2}\mu + \hat{u}_o = 0,\\
&v,\ \mu \geq 1,\ \hat{u}_o \geq 0.
\end{aligned}$$

This also gives $v^* = u^* = \hat{u}_o^* = 1$ so the applicable condition is (ii) in Theorem 2.1. Thus, RTS are decreasing at E', the point on the BCC efficiency frontier which is shown in Fig. 2.2.

2.6 Multiplicative Models

The treatments to this point have been confined to "qualitative" characterizations in the form of identifying whether RTS are "increasing," "decreasing," or "constant." There is a literature – albeit a relatively small one – which is directed to "quantitative" estimates of RTS in DEA. Examples are the treatment of scale elasticities in Banker et al. (1984), Førsund (1996) and Banker and Thrall (1992). However, there are problems in using the standard DEA models, as is done in these studies, to obtain scale elasticity estimates. Førsund (1996), for instance, lists a number of such problems. Also the elasticity values in Banker and Thrall (1992) are determined only within upper and lower bounds. This is an inherent limitation that arises from the piecewise linear character of the frontiers for these models. Finally, attempts to extend the Färe, Grosskopf and Lovell (1985, 1994) approaches to the determination of scale elasticities have not been successful. See the criticisms in Førsund (1996, p. 296) and Fukuyama (2000, p. 105). (Multiple output–multiple input production and cost functions, which meet the sub- and superadditivity requirements in economics, are dealt with in Panzar and Willig (1977). See also Baumol et al. (1982).)

This does not, however, exhaust the possibilities. There is yet another class of models referred to as "multiplicative models," which were introduced by this name into the DEA literature in Charnes et al. (1982) – see also Banker et al. (1981) – and extended in Charnes et al. (1983) to accord these models nondimensional (=units invariance) properties such as those we have just discussed. Although not used very much in applications these multiplicative models can provide advantages for extending the range of potential uses for DEA. For instance, they are not confined to efficiency frontiers that are concave. They can be formulated to allow the efficiency frontiers to be concave in some regions and nonconcave elsewhere. See Banker and Maindiratta (1986). They can also be used to obtain "exact" estimates of elasticities in manners that we now describe.

The models we use for this discussion are due to Banker and Maindiratta (1986) – where analytical characterizations are supplied along with confirmation in controlled-experimentally designed simulation studies.

We depart from the preceding development and now use an output-oriented model, which has the advantage of placing this development in consonance with the one in Banker and Maindiratta (1986) – viz.,

$$
\begin{aligned}
&\max \gamma_o, \\
&\text{subject to} \\
&\prod_{j=1}^{n} x_{ij}^{\lambda_j} \leq x_{io}, \qquad i = 1, \ldots, m, \\
&\prod_{j=1}^{n} y_{rj}^{\lambda_j} \geq \gamma_o y_{ro}, \quad r = 1, \ldots, s, \\
&\sum_{j=1}^{n} \lambda_j = 1, \\
&\gamma_o, \lambda_j \geq 0.
\end{aligned} \tag{2.19}
$$

To convert these inequalities to equations we use

$$e^{s_i^-} \prod_{j=1}^{n} x_{ij}^{\lambda_j} = x_{io}, \qquad i = 1, \ldots, m$$

and

$$e^{-s_r^+} \prod_{j=1}^{n} y_{rj}^{\lambda_j} = \gamma_o y_{ro}, \quad r = 1, \ldots, s \tag{2.20}$$

and replace the objective in (2.19) with $\gamma e^{\varepsilon\left(\sum_{r=1}^{s} s_r^+ + \sum_{i=1}^{m} s_i^-\right)}$, where $s_i^-, s_r^+ \geq 0$ represent slacks. Employing (2.20) and taking logarithms we replace (2.19) with

$$\begin{aligned}
&\min -\tilde{\gamma}_o - \varepsilon\left(\sum_{r=1}^{s} s_r^+ + \sum_{i=1}^{m} s_i^-\right),\\
&\text{subject to}\\
&\tilde{x}_{io} = \sum_{j=1}^{n} \tilde{x}_{ij}\lambda_j + s_i^-, \qquad i = 1, \ldots, m,\\
&\tilde{\gamma}_o + \tilde{y}_{ro} = \sum_{j=1}^{n} \tilde{y}_{rj}\lambda_j - s_r^+, \quad r = 1, \ldots, s,\\
&1 = \sum_{j=1}^{n} \lambda_j,\\
&\lambda_j, s_r^+, s_i^- \geq 0, \quad \forall j, r, i,
\end{aligned} \tag{2.21}$$

where "~" denotes "logarithm" so the $\tilde{x}_{ij}$, $\tilde{y}_{rj}$ and the $\tilde{\gamma}_o$, $\tilde{x}_{io}$, $\tilde{y}_{ro}$ are in logarithmic units.

The dual to (2.21) is

$$\begin{aligned}
&\max \sum_{r=1}^{s} \beta_r \tilde{y}_{ro} - \sum_{i=1}^{m} \alpha_i \tilde{x}_{io} - \alpha_o,\\
&\text{subject to}\\
&\sum_{r=1}^{s} \beta_r \tilde{y}_{rj} - \sum_{i=1}^{m} \alpha_i \tilde{x}_{ij} - \alpha_o \leq 0, \quad j = 1, \ldots, n,\\
&\sum_{r=1}^{s} \beta_r = 1,\\
&\alpha_i \geq \varepsilon, \ \beta_r \geq \varepsilon; \quad \alpha_o \text{ free in sign.}
\end{aligned} \tag{2.22}$$

Using α_i^*, β_r^* and α_o^* for optimal values, $\sum_{r=1}^{s} \beta_r^* \tilde{y}_{ro} - \sum_{i=1}^{m} \alpha_i^* \tilde{x}_{io} - \alpha_o^* = 0$ represents a supporting hyperplane (in logarithmic coordinates) for DMU_o, where efficiency

is achieved. We may rewrite this log-linear supporting hyperplane in terms of the original input/output values:

$$\prod_{r=1}^{s} y_{ro}^{\beta_r^*} = \mathrm{e}^{\alpha_o^*} \prod_{i=1}^{m} x_{io}^{\alpha_i^*}. \tag{2.23}$$

Then, in the spirit of Banker and Thrall (1992), we introduce

Theorem 2.7. Multiplicative Model RTS,

(i) RTS are increasing if and only if $\sum \alpha_i^* > 1$ for all optimal solutions to (2.23).
(ii) RTS are decreasing if and only if $\sum \alpha_i^* < 1$ for all optimal solutions to (2.23).
(iii) RTS are constant if and only if $\sum \alpha_i^* = 1$ for some optimal solutions to (2.23).

To see what this means we revert to the discussion of (2.11) and introduce scalars a, b in (aX_o, bY_o). In conformance with (2.23) this means

$$\mathrm{e}^{\alpha_o^*} \prod_{i=1}^{m} (ax_{io})^{\alpha_i^*} = \prod_{r=1}^{s} (by_{ro})^{\beta_r^*}, \tag{2.24}$$

so that the thus altered inputs and outputs satisfy this extension of the usual Cobb–Douglas types of relations.

The problem now becomes: given an expansion $a > 1$, contraction $a < 1$, or neither, i.e., $a = 1$, for application to all inputs, what is the value of b that positions the solution in the supporting hyperplane at this point? The answer is given by the following

Theorem 2.8. If (aX_o, bY_o) lies in the supporting hyperplane then $b = a^{\sum_{i=1}^{m} \alpha_i^*}$.

Proof. This proof is adopted from Banker et al. (2004). Starting with the expression on the left in (2.25) we can write

$$\mathrm{e}^{\alpha_o^*} \prod_{i=1}^{m} (ax_{io})^{\alpha_i^*} = a^{\sum_{i=1}^{m} \alpha_i^*} \mathrm{e}^{\alpha_o^*} \prod_{i=1}^{m} x_{io}^{\alpha_i^*} = \frac{a^{\sum_{i=1}^{m} \alpha_i^*}}{b} \prod_{r=1}^{s} (by_{ro})^{\beta_r^*}, \tag{2.25}$$

by using the fact that $\sum_{r=1}^{s} \beta_r^* = 1$ in (2.22) and $\mathrm{e}^{\alpha_o^*} \prod_{i=1}^{m} x_{io}^{\alpha_i^*} = \prod_{r=1}^{s} y_{ro}^{\beta_r^*}$ in (2.23).

Thus, to satisfy the relation (2.24) we must have $b = a^{\sum_{i=1}^{m} \alpha_i^*}$ as the theorem asserts. ■

Via this Theorem, we have the promised insight into reasons why more than proportionate output increases are associated with $\sum_{i=1}^{m} \alpha_i^* > 1$, less than proportionate increases are associated with $\sum_{i=1}^{m} \alpha_i^* < 1$ and CRS is the applicable condition when $\sum_{i=1}^{m} \alpha_i^* = 1$.

There may be alternative optimal solutions for (2.22) so the values for the α_i^* components need not be unique. For dealing with alternate optima, we return to (2.19) and note that a necessary condition for efficiency is $\gamma_o^* = 1$.

For full efficiency we must also have all slacks at zero in (2.20). An adaptation of (2.3) to the present problem, therefore, gives the following:

$$\prod_{j=1}^{n} x_{ij}^{\lambda_j^*} = e^{-s_i^{-*}} x_{io} = x'_{io}, \qquad i = 1, \ldots, m,$$
$$\prod_{j=1}^{n} y_{rj}^{\lambda_j^*} = e^{s_{ri}^{+*}} \gamma_o^* y_{ro} = y'_{ro}, \qquad r = 1, \ldots, s \tag{2.26}$$

and x'_{io}, y'_{ro} are the coordinates of the point on the efficiency frontier used to evaluate DMU_o.

Thus, we can extend the preceding models in a manner that is now familiar. Suppose that we have obtained an optimal solution for (2.22) with $\sum_{i=1}^{m} \alpha_i^* < 1$. We then utilize (2.26) to form the following problem

$$\begin{aligned}
&\max \sum_{i=1}^{m} \alpha_i, \\
&\text{subject to} \\
&\sum_{r=1}^{s} \beta_r \tilde{y}_{rj} - \sum_{i=1}^{m} \alpha_i \tilde{x}_{ij} - \alpha_o \leq 0, \quad j = 1, \ldots, n;\ j \neq 0, \\
&\sum_{r=1}^{s} \beta_r \tilde{y}'_{ro} - \sum_{i=1}^{m} \alpha_i \tilde{x}'_{io} - \alpha_o = 0, \\
&\sum_{r=1}^{s} \beta_r = 1, \\
&\sum_{i=1}^{m} \alpha_i \leq 1, \\
&\alpha_i \geq \varepsilon, \beta_r \geq \varepsilon; \quad \alpha_o \text{ free in sign.}
\end{aligned} \tag{2.27}$$

If $\sum_{i=1}^{m} \alpha_i^* = 1$ in (2.27), then RTS are constant by (iii) of Theorem 2.8. If the maximum is achieved with $\sum_{i=1}^{m} \alpha_i^* < 1$, however, condition (ii) of Theorem 2.7 is applicable and RTS are decreasing at the point x'_{io}, y'_{ro}; $i = 1, \ldots, m$; $r = 1, \ldots, s$.

If we initially have $\sum_{i=1}^{m} \alpha_i^* > 1$ in (2.22), we replace $\sum_{i=1}^{m} \alpha_i^* \leq 1$ with $\sum_{i=1}^{m} \alpha_i^* \geq 1$ in (2.27) and also change the objective to minimize $\sum_{i=1}^{m} \alpha_i^*$. If the optimal value is greater than one, then (i) of Theorem 2.7 is applicable and the RTS are increasing. On the contrary, if we attain $\sum_{i=1}^{m} \alpha_i^* = 1$ then condition (iii) applies and RTS are constant.

Theorem 2.8 also allows us to derive pertinent scale elasticities in a straightforward manner. Thus, using the standard logarithmic derivative formulas for elasticities, we obtain the following:

$$\frac{\mathrm{d} \ln b}{\mathrm{d} \ln a} = \frac{a}{b} \frac{\mathrm{d}b}{\mathrm{d}a} = \sum_{i=1}^{m} \alpha_i^*. \tag{2.28}$$

Consisting of a sum of component elasticities, one for each input, this overall measure of elasticity is applicable to the value of the multiplicative expression with which DMU_o is associated.

The derivation in (2.28) holds only for points where this derivative exists. However, we can bypass this possible source of difficulty by noting that Theorem 2.8 allows us to obtain this elasticity estimate via

$$\frac{\ln b}{\ln a} = \sum_{i=1}^{m} \alpha_i^*. \tag{2.29}$$

Further, as discussed in Cooper et al. (1996), it is possible to extend these concepts to the case in which all of the components of Y_o are allowed to increase by at least the factor b. However, we cannot similarly treat the constant, a, as providing an upper bound for the inputs since mix alterations are not permitted in the treatment of RTS in economics. See Varian (1984, p. 20) for requirements of RTS characterizations in economics.

In conclusion, we turn to properties of units invariance for these multiplicative models. Thus, we note that $\sum_{i=1}^{m} \alpha_i^*$ is units invariant by virtue of the relation expressed in (2.28). The property of units invariance is also exhibited in (2.29) since a and b are both dimension free. Finally, we also have the following.

Theorem 2.9. The model given in (2.19) and (2.20) is dimension free. That is, changes in the units used to express the input quantities x_{ij} or the output quantities y_{rj} in (2.19) will not affect the solution set or alter the value of $\max \gamma_o = \gamma_o^*$.

Proof. Let

$$\begin{aligned} x'_{ij} &= c_i x_{ij}, \quad x'_{io} = c_i x_{io}, \quad i = 1, \ldots, m, \\ y'_{rj} &= k_r y_{rj}, y'_{ro} = k_r y_{ro}, \quad r = 1, \ldots, s, \end{aligned} \tag{2.30}$$

where the c_i and k_r are any collection of positive constants. By substitution in the constraints for (2.20), we then have

$$\begin{aligned} e^{s_i^-} \prod_{j=1}^{n} x'_{ij} \lambda_j &= x'_{io}, \quad i = 1, \ldots, m, \\ e^{s_r^+} \prod_{j=1}^{n} y'_{rj} \lambda_j &= \gamma_o y'_{ro}, \quad r = 1, \ldots, s, \\ \sum_{j=1}^{n} \lambda_j &= 1, \ \lambda_j \geq 0, \quad j = 1, \ldots, n. \end{aligned} \tag{2.31}$$

Utilization of (2.30), therefore, gives

$$
\begin{aligned}
e^{s_i^-} c_i^{\sum_{j=1}^n \lambda_j} \prod_{j=1}^n x_{ij}^{\lambda_j} &= c_i x_{io}, && i = 1, \ldots, m, \\
e^{-s_r^+} k_r^{\sum_{j=1}^n \lambda_j} \prod_{j=1}^n y_{rj}^{\lambda_j} &= \gamma_o k_r y_{ro}, && r = 1, \ldots, s, \\
\sum_{j=1}^n \lambda_j &= 1, \quad \lambda_j \geq 0, && j = 1, \ldots, n.
\end{aligned} \tag{2.32}
$$

However, $\sum_{j=1}^n \lambda_j = 1$, so $c_i^{\sum_{j=1}^n \lambda_j} = c_i$ and $k_r^{\sum_{j=1}^n \lambda_j} = k_r \quad \forall i, r$. Therefore, these constants, which appear on the right and left of (2.32), all cancel. Thus, all solutions to (2.31) are also solutions to (2.20) and vice versa. It follows that the optimal value of one program is also optimal for the other. ■

We now conclude our discussion of these multiplicative models with the following:

Corollary to Theorem 2.9. The restatement of (2.20) in logarithmic form yields a model that is translation invariant.

Proof. Restating (2.31) in logarithmic form gives

$$
\begin{aligned}
s_i^- + \sum_{j=1}^n \left(\tilde{x}_{ij} + \tilde{c}_i\right) \lambda_j &= \tilde{x}_{io} + \tilde{c}_i, && i = 1, \ldots, m, \\
-s_r^+ + \sum_{j=1}^n \left(\tilde{y}_{rj} + \tilde{k}_r\right) \lambda_j &= \tilde{y}_{ro} + \tilde{k}_r + \tilde{\gamma}_o, && r = 1, \ldots, s, \\
\sum_{j=1}^n \lambda_j &= 1, \ \lambda_j \geq 0, && j = 1, \ldots, n.
\end{aligned} \tag{2.33}
$$

Once more utilizing $\sum \lambda_j = 1$ we eliminate the $\tilde{c}_i$ and $\tilde{k}_r$ on both sides of these expressions and obtain the same constraints as in (2.21). Thus, as before, the solution sets are the same and an optimum solution for one program is also optimal for the other – including the slacks. ■

2.7 Summary and Conclusion

Although we have now covered all of the presently available models, we have not covered all of the orientations in each case. Except for the multiplicative models, we have not covered output-oriented objectives for a variety of reasons. There are no real problems with the mathematical development, but further attention must be

devoted to how changes in input scale and input mix should be treated when all outputs are to be scaled up in the same proportions. See the discussion in Cooper et al. (1996).

As also noted in Cooper et al. (1996), the case of IRS can be clarified by using Banker's most productive scale size to write $(X_o\alpha,\ Y_o\beta)$. The case $1 < \beta/\alpha$ means that all outputs are increased by at least the factor β and RTS are increasing as long as this condition holds. The case $1 > \beta/\alpha$ has the opposite meaning – viz., no output is increasing at a rate that exceeds the rate at which all inputs are increased. Only for CRS do we have $1 = \beta/\alpha$, in which case all outputs and all inputs are required to be increasing (or decreasing) at the same rate so no mix change is involved for the inputs.

The results in this chapter (as in the literature to date) are restricted to this class of cases. This leaves unattended a wide class of cases. One example involves the case where management interest is centered on only subsets of the outputs and inputs. A direct way to deal with this situation is to partition the inputs and outputs of interest and designate the conditions to be considered by $(X_o^I\alpha, X_o^N, Y_o^I\beta, Y_o^N)$ where I designates the inputs and outputs that are of interest to management and N designates those which are not of interest (for such scale returns studies). Proceeding as described in the present chapter and treating X_o^N and Y_o^N as "exogenously fixed," in the spirit of Banker and Morey (1986), would make it possible to determine the situation for RTS with respect to the thus designated subsets. Other cases involve treatments with unit costs and prices as in FGL (1994) and Sueyoshi (1999).

The developments covered in this chapter have been confined to technical aspects of production. Our discussions follow a long-standing tradition in economics that distinguishes scale from mix changes by not allowing the latter to vary when scale changes are being considered. This permits the latter (i.e., scale changes) to be represented by a single scalar – hence the name. However, this can be far from actual practice, where scale and mix are likely to be varied simultaneously when determining the size and scope of an operation. See the comments by a steel industry consultant that are quoted in Cooper, Seiford, and Tone (2000, p. 130) on the need for reformulating this separation between mix and scale changes to achieve results that more closely conform to needs and opportunities for use in actual practice.

There are, of course, many other aspects to be considered in treating RTS besides those attended to in the present chapter. Management efforts to maximize profits, even under conditions of certainty, require simultaneous determination of scale, scope, and mix magnitudes with prices and costs known, as well as the achievement of the technical efficiency, which is always to be achieved with *any* set of positive prices and costs. The topics treated in this chapter do not deal with such price–cost information. Moreover, the focus is on ex post facto analysis of already effected decisions. This can have many uses, especially in the control aspects of management where evaluations of performance are required. Left unattended in this chapter, and in much of the DEA literature, is the ex ante (planning) problem of how to use this knowledge to determine how to blend scale and scope with mix and other efficiency considerations when effecting future-oriented decisions.

Appendix

In this Appendix, we first present the FGL approach. We then present a simple RTS approach without the need for checking the multiple optimal solutions as in Zhu and Shen (1995) and Seiford and Zhu (1999) where only the BCC and CCR models are involved. This approach will substantially reduce the computational burden because it relies on the standard CCR and BCC computational codes (see Zhu (2009) for a detailed discussion).

To start, we add to the BCC and CCR models by the following DEA model whose frontier exhibits nonincreasing returns to scale (NIRS), as in Färe, Grosskopf and Lovell (FGL 1985, 1994)

$$\begin{aligned}
&\theta^*_{\text{NIRS}} = \min \theta_{\text{NIRS}},\\
&\text{subject to}\\
&\theta_{\text{NIRS}} x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-, \quad i = 1, 2, \ldots, m,\\
&y_{ro} \qquad\quad = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+, \quad r = 1, 2, \ldots, s,\\
&1 \geq \sum_{j=1}^{n} \lambda_j,\\
&0 \leq \lambda_j, s_i^-, s_r^+ \quad \forall\, i, r, j.
\end{aligned} \tag{2.34}$$

The development used by FGL (1985, 1994) rests on the following relation

$$\theta^*_{\text{CCR}} \leq \theta^*_{\text{NIRS}} \leq \theta^*_{\text{BCC}},$$

where "*" refers to an optimal value and θ^*_{NIRS} is defined in (2.34), while θ^*_{BCC} and θ^*_{CCR} refer to the BCC and CCR models as developed in Theorems 2.3 and 2.4.

FGL utilize this relation to form ratios that provide measures of RTS. However, we turn to the following tabulation that relates their RTS characterization to Theorems 2.3 and 2.4 (and accompanying discussion). See also Färe and Grosskopf (1994), Banker et al. (1996b), and Seiford and Zhu (1999)

	FGL Model	RTS	CCR Model
Case 1	If $\theta^*_{\text{CCR}} = \theta^*_{\text{BCC}}$	Constant	$\sum \lambda_j^* = 1$
Case 2	If $\theta^*_{\text{CCR}} < \theta^*_{\text{BCC}}$ then		
Case 2a	If $\theta^*_{\text{CCR}} = \theta^*_{\text{NIRS}}$	Increasing	$\sum \lambda_j^* < 1$
Case 2b	If $\theta^*_{\text{CCR}} < \theta^*_{\text{NIRS}}$	Decreasing	$\sum \lambda_j^* > 1$

It should be noted that the problem of nonuniqueness of results in the presence of alternative optima is not encountered in the FGL approach (unless output-oriented as well as input-oriented models are used), whereas they do need to be coincided, as in Theorem 2.3. However, Zhu and Shen (1995) and Seiford and Zhu (1999) develop an alternative approach that is not troubled by the possibility of such alternative optima.

We here present their results with respect to Theorems 2.3 and 2.4 (and accompanying discussion). See also Zhu (2009).

	Seiford and Zhu (1999)	RTS	CCR Model
Case 1	If $\theta^*_{\text{CCR}} = \theta^*_{\text{BCC}}$	Constant	$\sum \lambda^*_j = 1$
Case 2	$\theta^*_{\text{CCR}} \neq \theta^*_{\text{BCC}}$		
Case 2a	If $\sum \lambda^*_j < 1$ in any CCR outcome	Increasing	$\sum \lambda^*_j < 1$
Case 2b	If $\sum \lambda^*_j > 1$ in any CCR outcome	Decreasing	$\sum \lambda^*_j > 1$

The significance of Seiford and Zhu's (1999) approach lies in the fact that the possible alternate optimal λ^*_j obtained from the CCR model only affect the estimation of RTS for those DMUs that truly exhibit CRS and have nothing to do with the RTS estimation on those DMUs that truly exhibit IRS or DRS. That is, if a DMU exhibits IRS (or DRS), then $\sum_j^n \lambda^*_j$ must be less (or greater) than one, no matter whether there exist alternate optima of λ_j, because these DMUs do not lie in the MPSS region. This finding is also true for the u^*_o obtained from the BCC multiplier models.

Thus, in empirical applications, we can explore RTS in two steps. First, select all the DMUs that have the same CCR and BCC efficiency scores regardless of the value of $\sum_j^n \lambda^*_j$ obtained from model (2.5). These DMUs are CRS. Next, use the value of $\sum_j^n \lambda^*_j$ (in any CCR model outcome) to determine the RTS for the remaining DMUs. We observe that in this process we can safely ignore possible multiple optimal solutions of λ_j.

References

Banker RD. Estimating most productive scale size using Data Envelopment Analysis. Eur J Oper Res. 1984;17:35–44.

Banker RD, Bardhan I, Cooper WW. A note on returns to scale in DEA. Eur J Oper Res. 1996a;88:583–5.

Banker RD, Chang H, Cooper WW. Equivalence and implementation of alternative methods for determining returns to scale in Data Envelopment Analysis. Eur J Oper Res. 1996b;89:473–81.

Banker R, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manag Sci. 1984;30:1078–92.

Banker RD, Charnes A, Cooper WW, Schinnar A. A bi-extremal principle for Frontier Estimation and Efficiency Evaluation. Manag Sci. 1981;27:1370–82.

Banker RD, Maindiratta A. Piecewise loglinear estimation of efficient production surfaces. Manag Sci. 1986;32:126–35.

Banker RD, Morey R. Efficiency analysis for exogenously fixed inputs and outputs. Oper Res. 1986;34:513–21.
Banker RD, Thrall RM. Estimation of returns to scale using Data Envelopment Analysis. Eur J Oper Res. 1992;62:74–84.
Banker, R.D., Cooper, W.W., Seiford, L.M., Thrall, R.M. and Zhu, J, Returns to scale in different DEA models, European Journal of Operational Research, 2004;154:345–362.
Baumol WJ, Panzar JC, Willig RD. Contestable markets. New York: Harcourt Brace Jovanovich; 1982.
Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res. 1978;2:429–44.
Charnes A, Cooper WW, Seiford LM, Stutz J. A multiplicative model for efficiency analysis. Socioecon Plann Sci. 1982;16:213–24.
Charnes A, Cooper WW, Seiford LM, Stutz J. Invariant multiplicative efficiency and piecewise Cobb-Douglas envelopments. Oper Res Lett. 1983;2:101–3.
Cooper WW, Park KS, Pastor JT. RAM: A range adjusted measure of efficiency. J Product Anal. 1999;11:5–42.
Cooper WW, Seiford LM, Tone K. Data envelopment analysis: a comprehensive text with models, references and DEA-Solver Software applications. Boston: Kluwer Academic Publishers; 2000.
Cooper WW, Thompson RG, Thrall RM. Extensions and new developments in DEA. Ann Oper Res. 1996;66:3–45.
Färe R, Grosskopf S. Estimation of returns to scale using data envelopment analysis: a comment. Eur J Oper Res. 1994;79:379–82.
Färe R, Grosskopf S, Lovell CAK. The measurement of efficiency of production. Boston: Kluwer Nijhoff Publishing; 1985.
Färe R, Grosskopf S, Lovell CAK. Production frontiers. Cambridge: Cambridge University Press; 1994.
Førsund FR. On the calculation of scale elasticities in DEA models. J Product Anal. 1996;7:283–302.
Frisch RA. Theory of production. Dordrecht: D. Rieoel; 1964.
Fukuyama H. Returns to scale and scale elasticity in Data Envelopment Analysis. Eur J Oper Res. 2000;125:93–112.
Golany B, Yu G. Estimating returns to scale in DEA. Eur J Oper Res. 1994;103:28–37.
Panzar JC, Willig RD. Economies of scale in multi-output production. Q J Econ. 1977; XLI:481–493.
Seiford LM, Zhu J. An investigation of returns to scale under Data Envelopment Analysis. Omega. 1999;27:1–11.
Sueyoshi T. DEA duality on returns to scale (RTS) in production and cost analyses: an occurrence of multiple solutions and differences between production-based and cost-based RTS estimates. Manag Sci. 1999;45:1593–608.
Thrall RM. Duality, classification and slacks in DEA. Ann Oper Res. 1996a;66:109–38.
Thrall RM. The lack of invariance of optimal dual solutions under translation invariance. Ann Oper Res. 1996b;66:103–8.
Varian H. Microeconomic analysis. New York: W.W. Norton; 1984.
Zhu J. Setting scale efficient targets in DEA via returns to scale estimation methods. J Oper Res Soc. 2000;51(3):376–8.
Zhu J. Quantitative models for performance evaluation and benchmarking: data envelopment analysis with spreadsheets. 2nd ed. Boston: Springer Science; 2009.
Zhu J, Shen Z. A discussion of testing DMUs' returns to scale. Eur J Oper Res. 1995;81:590–6.

Chapter 3
Sensitivity Analysis in DEA*

William W. Cooper, Shanling Li, Lawrence M. Seiford, and Joe Zhu

Abstract This chapter presents some of the recently developed analytical methods for studying the sensitivity of DEA results to variations in the data. The focus is on the stability of classification of DMUs (decision making units) into efficient and inefficient performers. Early work on this topic concentrated on developing algorithms for conducting such analyses after it was noted that standard approaches for conducting sensitivity analyses in linear programming could not be used in DEA. However, recent work has bypassed the need for such algorithms. It has also evolved from the early work that was confined to studying data variations in one input or output for one DMU. The newer methods described in this chapter make it possible to analyze the sensitivity of results when all data are varied simultaneously for all DMUs.

Keywords Data envelopment analysis • Efficiency • Stability • Sensitivity

3.1 Introduction

This chapter surveys analytical approaches that have been developed to treat sensitivity and stability analyses in data envelopment analysis (DEA). We may classify the approaches to this topic into two categories that we can characterize as (a) substantive and (b) methodological. The term "substantive" refers to generalizations or characterizations directed to the properties of DEA. An early

*Part of the material in this chapter is adapted from article in the Journal of Productivity Analysis, Cooper WW, Li S, Seiford LM, Tone K, Thrall RM, Zhu J, Sensitivity and stability analysis in DEA: some recent developments, 2001, 15, 217–46, and is published with permission from Kluwer Academic Publishers.

J. Zhu (✉)
School of Business, Worcester Polytechnic Institute, Worcester, MA 01609, USA
e-mail: jzhu@wpi.edu

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_3,

example is *Measuring Efficiency: An Assessment of Data Envelopment Analysis. New Directions for Program Evaluation*. Authored by Sexton et al. (1986), this book concludes that DEA results are likely to be unstable because its evaluations are based on "outlier" observations. We may contrast this with the "methodological" approaches to sensitivity and stability analysis. As the term suggests, this category is concerned with the development of tools and concepts that can be used to determine the degree of sensitivity to data variations in any particular application of DEA.

Possibly because they were not formulated in a manner that could provide guidance as to where further research or adaptations for use might best be pointed, little progress has been made in such substantive approaches. By way of contrast, there has been a great deal of progress in the category that we have referred to as methodological approaches. The early work in this area, which focused on analyses of a single input or output for a single decision making unit (DMU), has now moved to sensitivity analyses directed to evaluating the stability of DEA results when *all* inputs and outputs are varied simultaneously in *all* DMUs. Unlike other, looser characterizations of stability (as in the substantive approaches), these analytical approaches have been precise as to the nature of the problems to be treated. They have focused almost exclusively on changes from efficient to inefficient status for the DMUs being analyzed.

This chapter is devoted to these methodological approaches. Hence it, too, focuses on the sensitivity and stability of efficient vs. inefficient classifications of DMUs under different ranges of data variations.

The approaches to be examined are deterministic in the same sense that the models covered in the preceding chapters in this handbook are deterministic and so we leave the use of stochastic approaches with accompanying tests of stability and significance to later chapters in this handbook, as in the chapter by Banker this volume. Attention here is confined to studies involving only variations in data. Other topics such as sensitivity to model changes or diminution and augmentation in the number of DMUs (as in the statistical studies of sampling distributions) are covered in the chapter of this handbook by Simar and Wilson this volume.

3.2 Sensitivity Analysis Approaches

The topic of sensitivity (=stability or robustness) analysis has taken a variety of forms in the DEA literature. One part of this literature studies responses with given data when DMUs are deleted or added to the set being considered. See Wilson (1995) and also see the discussion of "window analysis" in Chap. 1 of this handbook. Another part of the literature deals with increases or decreases in the number of inputs and outputs to be treated. Analytically oriented treatments of these topics are not lacking but most of the literature has taken the form of simulation studies, as in Banker et al. (1996). We also do not examine sensitivity studies related to choices of different DEA models as in Ahn and Seiford (1993).

In any case, we restrict the discussion in this chapter to analytically formulated (mathematical) methods for examining stability and sensitivity of results to data variations with given variables and given models.

As in statistics or other empirically oriented methodologies, there is a problem involving degrees of freedom, which is compounded in DEA because of its orientation to *relative* efficiency. In the envelopment model, the number of degrees of freedom will increase with the number of DMUs and decrease with the number of inputs and outputs. A rough rule of thumb which can provide guidance is to choose a value of n that satisfies $n \geq \max\{m \times s, 3(m+s)\}$, where $n =$ number of DMUs, $m =$ number of inputs, and $s =$ number of outputs. Hereafter, we assume that this (or other) degrees of freedom conditions are satisfied and that there is no trouble from this quarter.

3.2.1 Algorithmic Approaches

Research on analytical approaches to sensitivity analysis in DEA was initiated in Charnes et al. (1985) where it was noted that the methods used for this purpose in linear programming were not appropriate to DEA. The discussion in Charnes et al. (1985) took the form of algorithmic developments which built on the earlier work in Charnes and Cooper (1968) – after noting that variations in the data for the DMU_o being analyzed could alter the inverse matrix that is generally used in linear programming approaches to sensitivity analyses. Building on the earlier work of Charnes and Cooper (1968), this work by Charnes et al. (1985) was therefore directed to developing algorithms that would avoid the need for additional matrix inversions. Originally confined to treating a single input or output this line of work was extended and improved in a series of papers published by Charnes and Neralic. For a summary discussion, see Charnes and Neralic (1992) and Neralic (1997).

3.2.2 Metric Approaches

Another avenue for sensitivity analysis opened by Charnes et al. (1992) bypasses the need for these kinds of algorithmic forays by turning to metric concepts. The basic idea is to use concepts such as "distance" or "length" (=norm of a vector) in order to determine "radii of stability" within which the occurrence of data variations will not alter a DMU's classification from efficient to inefficient status (or vice versa).

The resulting classifications obtained from these "radii of stability" can range from "unstable" to "stable" with the latter being identified by a radius of some finite value within which no reclassification will occur. For example, a point like F in Fig. 3.1 is identified as stable. A point like A, however, is unstable because

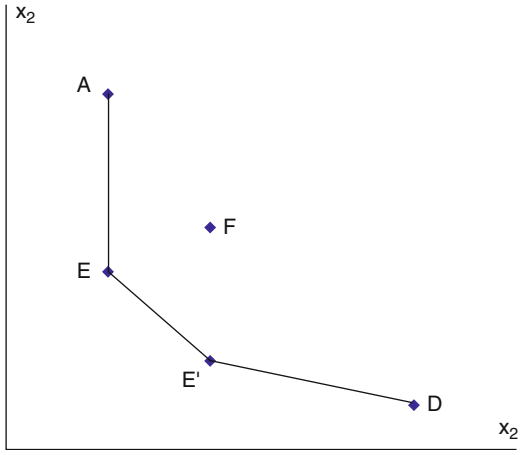


Fig. 3.1 Stable and unstable DMUs

an infinitesimal perturbation to the right or above its present position would alter its status from efficient to inefficient.

A variety of metrics and models are examined by Charnes and Cooper (1961). Here, however, attention will be confined to the Chebychev ($=l_\infty$) norm, as in the following model taken from Charnes and Neralic (1992, p. 795),

$$\begin{aligned}
&\max \delta \\
&\text{subject to} \\
&y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ - \delta d_r^+, \quad r = 1, \ldots, s \\
&x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- + \delta d_i^-, \quad i = 1, \ldots, m \\
&1 = \sum_{j=1}^{n} \lambda_{\mathrm{j}}
\end{aligned} \tag{3.1}$$

with all variables (including δ) constrained to be nonnegative while the d_r^+ and d_i^- are fixed constants (or weights) which we now equate to unity. For instance, with all $d_i^- = d_r^+ = 1$ the solution to (3.1) may be written as

$$\begin{aligned}
&\sum y_{rj}\lambda_j^* - s_r^{+*} = y_{ro} + \delta^*, \quad r = 1, \ldots, s \\
&\sum_{j=1}^{n} x_{ij}\lambda_j^* + s_i^{-*} = x_{io} - \delta^*, \quad i = 1, \ldots, m
\end{aligned} \tag{3.2}$$

where "*" indicates an optimum value and the value of δ* represents the maximum that this metric allows consistent with the solution on the left.

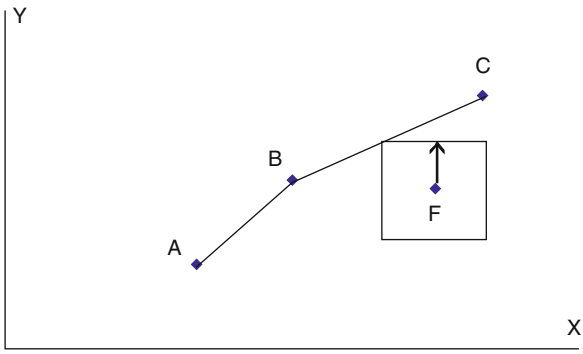


Fig. 3.2 A radius of stability

The formulation in (3.2) is for an inefficient DMU which continues to be inefficient for all data alterations from y_{ro} to $y_{ro} + \delta_r^*$ and from x_{io} to $x_{io} - \delta^*$. This is intended to mean that no reclassification to efficient status will occur within the open set defined by the value of $0 \leq \delta^*$, which is referred to as a "radius of stability." See, for example, the point F in Fig. 3.2 which is centered in the square (or box) which is referred to as a "unit ball" defined by this C (=Chebyshev) norm.

The above model dealt with improvements in *both* inputs and outputs that could occur for an *inefficient* point before its status would change to efficient – as in the upper left-hand corner of the square surrounding F in Fig. 3.2. The treatment of *efficient* points proceeds in the direction of "worsening" outputs and inputs as in the following model:

$$\begin{aligned}
&\min \delta \\
&\text{subject to} \\
&y_{ro} = \sum_{j=1, j\neq 0}^{n} y_{rj}\lambda_j - s_r^+ + \delta, \quad r = 1, \ldots, s \\
&x_{io} = \sum_{j=1, j\neq 0}^{n} x_{ij}\lambda_j + s_i^- - \delta, \quad i = 1, \ldots, m \\
&1 = \sum_{j=1, j\neq 0}^{n} \lambda_j
\end{aligned} \tag{3.3}$$

where, again, all variables are constrained to be nonnegative.

In this case $j \neq o$ refers to the efficient DMU_o that is being analyzed. Removal of this DMU_o is necessary because the solution will otherwise be $\delta^* = 0$ with $\lambda_o^* = 1$ and thereby indicate that the efficient DMU_o has a zero radius of stability.

We can generalize Charnes et al.'s (1992) instability property as follows.

Definition 3.1. The coordinates of the point associated with an efficient DMU will always have both efficient and inefficient points within a radius of $\varepsilon > 0$, however, small the value of ε.

Definition 3.2. Any point which has this property is unstable.

We can illustrate its applicability by noting that point A, which is not efficient in Fig. 3.1, has this property, since a slight variation to the left will change its status from inefficient to efficient. In any case, a solution, δ^*, provides a radius in the Chebyshev norm that is to be attained before an efficient DMU is changed to inefficient status. We illustrate this with the following example.

Example. To portray what is happening, we synthesize an example from Fig. 3.2 by assigning coordinates to these points in the order (x, y) as follows: $A = (1,1)$, $B = (3,4)$, and $C = (4,5)$. To simplify matters we remove F and consider only the efficient points A and C to determine a radius of stability for B. This is accomplished by applying (3.3) as follows:

$$\begin{aligned} &\min \delta \\ &\text{subject to} \\ &4 = 1\lambda_A + 5\lambda_C - s^+ + \delta \\ &3 = 1\lambda_A + 4\lambda_C + s^- - \delta \\ &1 = \lambda_A + \lambda_C. \end{aligned}$$

With all variables constrained to be nonnegative the values $\lambda_A^* = 0.29$, $\lambda_C^* = 0.71$, and $\delta^* = 0.15$ with both slacks zero optimally satisfy all constraints.

With $\delta^* = 0.15$ as the radius of stability we have $4 - \delta^* = 3.85$ as the worsened output and $3 + \delta^* = 3.15$ as the worsened input values: the thus altered values of B remain efficient if no further worsenings occur. To confirm this latter point we may note that the line segment connecting A and C represents the efficient frontier for the reduced convex set that remains after B is removed. This segment is on the line $y = -1/3 + 4/3x$. Substituting $3 + \delta^* = x = 3.15$ in this expression yields $y = 3.85 = 4 - \delta^*$. Hence, these worsened values of B place it on the efficient frontier of this reduced convex set where it retains its efficient status. Thus, as is evident in this example, the model (3.3) seeks to minimize the value of the norm that brings the thus removed point into contact with the frontier of the reduced convex set.

The above formulations are recommended by Charnes et al. (1992) only as the start for a sensitivity analysis. Because, inter alia, this (Chebychev) norm does not reflect nonzero slacks which may be present. However, Charnes et al. (1992) provide other models, such as additive model formulations which utilize the ℓ_1 metric, account for all such inefficiencies.

3.2.3 Multiplier Model Approaches

The two approaches described above – i.e., the algorithmic and metric approaches – use DEA "envelopment models" to treat one DMU at a time. Extensions are needed if we are to treat cases where the DMUs are numerous and where it is not clear which ones require attention. Ideally, it should be possible to vary all data simultaneously for all DMUs until the status of at least one DMU is changed from inefficient to efficient or vice versa. An approach initiated in Thompson et al. (1994) moves in this direction in a manner that we now describe.

For this purpose, we record the following dual pair of problems from Thompson et al. (1996):

Envelopment model	Multiplier model
$\text{minimize}_{\theta,\lambda}\ \theta$	$\text{maximize}_{u,v}\ z = uy_o$
subject to	subject to
$Y\lambda \geq y_o$	$u \geq 0$
$\theta x_o - X\lambda \geq 0$	$v \geq 0$
$\lambda \geq 0$	$uY - vX \leq 0$
θ unrestricted	$x_o = 1$

(3.4)

where Y, X, and y_o, x_o are data matrices and vectors of outputs and inputs, respectively, and λ, u, v are vectors of variables (λ: a column vector; u and v: row vectors). θ, a scalar, which can be positive, negative, or zero in the envelopment model, is the source of the condition $vx_o = 1$ which appears at the bottom of the multiplier model.

The above arrangement, we might note systematically relates the elements of these dual problems to each other in the following fashion. Nonnegativity for the vectors u and v on the right in the multiplier model is associated with the inequality constraints on their left in the envelopment model. Similarly, nonnegativity for the vector λ on the left is associated with the inequality conditions on the right. Finally, the unrestricted variable θ in the envelopment model is associated with the condition $vx_o = 1$ on its right.

We now observe that no allowance for nonzero slacks is made in the objective of the above envelopment model. Hence, the variables in the multiplier model are constrained only to be nonnegative. That is, the *positivity* requirement associated with the non-Archimedean element, ε described in Chap. 1, is absent from both members of this dual pair. For present purposes, however, we only need to note that the sensitivity analyses we will now be considering are centered around the set of efficient extreme points and these points always have a unique optimum with nonzero slack solutions for the envelopment (but not the multiplier) model (see Charnes et al. 1991).

The analysis used by Thompson et al. (1994) is carried forward via the multiplier models. This makes it possible to exploit the fact that the values u^*, v^* which are optimal for the DMU being evaluated will remain valid over some (generally positive) range of variation in the data.

Following Thompson et al., we exploit this property by defining a new vector $w = (u, v)$ which we use to define a function $h_j(w)$ as follows:

$$h_j(w) = \frac{f_j(w)}{g_j(w)} = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}}. \tag{3.5}$$

Next, let

$$h_o(w) = \max_{j=1,\ldots,n} h_j(w) \tag{3.6}$$

so that

$$h_o(w) \geq h_j(w), \quad \forall j. \tag{3.7}$$

It is now to be noted that (3.5) returns matters to the nonlinear version of the CCR ratio form given by (1.1) and (IR) in Chap. 1. Hence, we need not be concerned with continued satisfaction of the norm condition, $vx_o = 1$ in (3.4) when we study variations in the data in the manner now to be described.

When an optimal w^* does not satisfy (3.7), the DMU$_o$ being evaluated is said to be "radial inefficient." The term is appropriate because this means that $\theta^* < 1$ will occur in the envelopment model. The full panoply of relations between the CCR ratio, multiplier, and envelopment models is thus brought into play without any need for extensive computations or analyses.

Among the frontier points for which $\theta^* = 1$, attention is directed by Thompson et al. to "extreme efficient points." In particular, attention is centered on points in the set E for which, for some multiplier w^*,

$$h_o(w^*) > h_j(w^*), \quad \forall j \neq 0. \tag{3.8}$$

This (strict) inequality will generally remain valid over some range of variation in the data. Hence, in more detail, we will have

$$h_o(w^*) = \frac{\sum_{r=1}^{s} u_r^* y_{ro}}{\sum_{i=1}^{m} v_i^* x_{io}} > \frac{\sum_{r=1}^{s} u_r^* y_{rj}}{\sum_{i=1}^{m} v_i^* x_{ij}} = h_j(w^*), \quad \forall j \neq o, \tag{3.9}$$

Table 3.1 Data for a sensitivity analysis

	E-efficient			Not efficient		
DMU	1	2	3	4	5	6
Output: y	1	1	1	1	1	1
Input: x1	4	2	1	2	3	4
Input: x2	1	2	4	3	2	4

E-efficient = extreme point efficient

Table 3.2 Initial solutions

	DMU1	DMU2	DMU3
DMU	$h_j(w^1)$	$h_j(w^2)$	$h_j(w^3)$
1	1.000	0.800	0.400
2	0.714	1.000	0.714
3	0.400	0.800	1.000
4	0.500	0.800	0.667
5	0.667	0.800	0.550
6	0.357	0.500	0.357

which means that DMU_o is more efficient than any other DMU_j and hence will be rated as fully efficient by DEA.

Thompson et al. (1994) employ a ranking principle which they formulate as:

> "If DMU_o is more efficient than all of the other DMU_j relative to the vector w^*, then DMU_o is said to be top ranked."

Thus, holding w^* fixed, the data are varied and DMU_o is then said to be "top ranked" as long as (3.8) continues to hold for the data variations under consideration.

Thompson et al. (1994) carry out experiments in which the data are allowed to vary in different ways – including allowing the data variations to occur at random. Among these possibilities, we examine only the following one: for DMU_o, which is extreme efficient, the outputs will all be decreased and the inputs will all be increased by a stipulated amount (or percentage). This same treatment is accorded to the other DMUs which are efficient. For the other DMU_j, which are all inefficient, the reverse adjustment is made: all outputs are increased and all inputs are decreased in these same amounts (or percentages). In this way, the value of the ratio will be decreased for DMU_o in (3.8) and in the other extreme efficient DMUs while the ratios for the other DMU_j will be increased. Continuing in this manner a reversal can be expected to occur at some point – in which case DMU_o will no longer be "top ranked" which means that it will then lose the status of being fully (DEA) efficient.

Table 3.1 taken from Thompson et al. (1994) will be used to illustrate the procedure in a simple manner by varying only the data for the inputs x_1, x_2 in this table.

To start the sensitivity analysis, Table 3.2 records the initial solutions obtained by applying the multiplier model in (3.4) to these data for each of DMU_1, DMU_2,

Table 3.3 5% increments and decrements

	DMU1	DMU2	DMU3
DMU	$h_j(w^1)$	$h_j(w^2)$	$h_j(w^3)$
1	0.952	0.762	0.381
2	0.680	0.952	0.680
3	0.381	0.762	0.952
4	0.526	0.842	0.702
5	0.702	0.842	0.552
6	0.376	0.526	0.376

and DMU_3 which are all E (=extreme point) efficient. As can be seen, these solutions show DMU_1, DMU_2, and DMU_3 to be top ranked in their respective columns.

The gaps between the top and other ranks from these results show that some range of data variation can be undertaken without changing this top-ranked status in any of these three columns. To start we therefore follow Thompson et al. and introduce 5% increase in each of x_1 and x_2 for DMU_1, DMU_2, and DMU_3. Simultaneously, we decrease these inputs by 5% for the other DMUs to obtain Table 3.3.

As can be seen in Table 3.3, each of DMU_1, DMU_2, and DMU_3 maintain their "top-ranked status" and hence continue to be DEA fully efficient (relatively). Nor is this the end of the line. Continuing in this 5% increment–decrement fashion, a 15% increment–decrement is needed, as Thompson et al. (1994) report, for a first displacement in which DMU_2 is replaced by DMU_4 and DMU_5. Continuing further, a 20% increment–decrement is needed to replace DMU_4 with DMU_1 and, finally, still further incrementing and decrementing is needed to replace DMU_4 with DMU_1 as top ranked.

Note that the $h_j(w)$ values for all of the *efficient* DMUs decrease in every column when going from Table 3.2 to Table 3.3 and, simultaneously, the $h_j(w)$ values increase for the *inefficient* DMUs. The same behavior occurs for the other data variations.

As noted in Thompson et al. (1994), this robust behavior is obtained for extreme efficient DMUs – which are identified by their satisfaction of the strong complementary slackness condition for which a gap will appear like ones between the top and second rank shown in every column of Table 3.2. In fact, the choice of w^* can affect the degree of robustness as reported in Thompson et al. (1996) where use of an interior point algorithm produces a w^* closer to the "analytic center" and this considerably increases the degree of robustness for the above example.

As just noted, Thompson et al. (1994) confine their analysis to points which are extreme efficient. They then utilize the "strong form of the complementary slackness principle" to ensure that all of the dual (=multiplier) values are positive – as in Table 3.4 – and this avoids possible troubles associated with the appearance of zeros in the ratios represented in (3.9).

Table 3.4 Optimal dual (=multiplier) values

	Outputs	Inputs	
	u^*	v_1^*	v_2^*
DMU_1	1.0	0.10	0.60
DMU_2	1.0	0.25	0.25
DMU_3	1.0	0.60	0.10

Source: Thompson et al. (1994)

To see what is happening with this approach we briefly review the principles involved. First, we recall the "complementary slackness principle" of linear programming which we represent in the following form:

$$\begin{aligned} s_r^{+*} u_r^* &= 0, \quad r = 1, \ldots, s \\ s_i^{-*} v_i^* &= 0, \quad i = 1, \ldots, m, \end{aligned} \tag{3.10}$$

where $s_r^{+*} \geq 0$ and $s_i^{-*} \geq 0$ are optimal output and input slacks, respectively, for the envelopment model, and u_r^* and v_i^* are the dual (=multiplier) values associated with the multiplier model in (3.4). Verbally, this means that at least one (and possibly both) variable in each pair must be zero in their corresponding optimal solutions.

The "strong principle of complementary slackness" may be represented

$$\begin{aligned} s_r^{+*} + u_r^* &> 0, \quad r = 1, \ldots, s \\ s_i^{-*} + v_i^* &> 0, \quad i = 1, \ldots, m. \end{aligned} \tag{3.11}$$

In short, the possibility of both variables being zero is eliminated.

By restricting attention to the set of efficient points, Thompson et al. insure that only the "multiplier" variables will be positive. This occurs because the solutions associated with the extreme efficient points in the envelopment model are always unique with all slacks at zero. Conformance with the conditions in (3.11), therefore, insures that all $s + m$ multiplier values will be positive and the results portrayed in Table 3.4 are a consequence and this property. Having all of these variables positive is referred to as a solution with "full dimensionality" by Thompson et al. (1994).

3.2.4 A Two-Stage Alternative

This use of strong complementary slackness involves recourse to special algorithms like the interior point methods developed in Thompson et al. (1996). To avoid this need we therefore suggest an alternative in the form of a two-stage approach as follows: stage one utilizes any of the currently available computer codes. This willyield satisfactory results in many (if not most) cases even though strong

Table 3.5 Multiplier values

	Outputs	Inputs	
	u^*	v_1^*	v_2^*
DMU_1	1	0.167	0.333
DMU_2	1	0.333	0.167
DMU_3	1	1	0

complementarity is not satisfied. Stage two is to be invoked only when these results are not satisfactory – in which case recourse to special algorithms will be needed like those described in Thompson et al. (1996).

We illustrate this approach by using the data of Table 3.1 to apply the Thompson et al. (1994) formulas in the manner that we described in the preceding section of this chapter. In this illustrative application, we obtain solutions to the multiplier models for DMU_1, DMU_2, DMU_3 which we display in Table 3.5. As can be seen, some of these multiplier values differ from the ones exhibited in Table 3.4. In addition, a zero value appears for v_2^* in the row for DMU_3. Nevertheless, the results do not differ greatly from those reported by Thompson et al. and, in particular, the same kind of very robust results are secured.

To illustrate the Thompson et al. procedure we develop this example in more detail. For this purpose, we focus on the multiplier values for DMU_1, which are $w^1 = (u^*, v_1^*, v_2^*) = (1, 0.167, 0.333)$ taken from Table 3.5. Applying these values to the data of Table 3.1 for each of these $j = 1, \ldots, 6$ DMUs we obtain

$$h_1(w^1) = \frac{u^*}{v_1^* x_{11} + v_2^* x_{21}} = \frac{u^*}{4v_1^* + 1v_2^*} = \frac{1}{0.667 + 0.333} = 1$$

$$h_2(w^1) = \frac{u^*}{v_1^* x_{12} + v_2^* x_{22}} = \frac{u^*}{2v_1^* + 2v_2^*} = \frac{1}{0.333 + 0.667} = 1$$

$$h_3(w^1) = \frac{u^*}{v_1^* x_{13} + v_2^* x_{23}} = \frac{u^*}{1v_1^* + 4v_2^*} = \frac{1}{0.167 + 1.333} = 0.667$$

$$h_4(w^1) = \frac{u^*}{v_1^* x_{14} + v_2^* x_{24}} = \frac{u^*}{2v_1^* + 3v_2^*} = \frac{1}{0.333 + 1.0} = 0.75$$

$$h_5(w^1) = \frac{u^*}{v_1^* x_{15} + v_2^* x_{25}} = \frac{u^*}{3v_1^* + 2v_2^*} = \frac{1}{0.5 + 0.667} = 0.857$$

$$h_6(w^1) = \frac{u^*}{v_1^* x_{16} + v_2^* x_{26}} = \frac{u^*}{4v_1^* + 4v_2^*} = \frac{1}{0.667 + 1.333} = 0.5.$$

In contrast to the results displayed in column 1 of Table 3.2, we find that DMU_2 has moved up to tie DMU_1 in top-rank position, since $h_1(w^1) = h_2(w^1) = 1$. Nevertheless, DMU_1 is not displaced. Since DMU_1 continues to maintain its efficient status, we consider the result to be satisfactory even though the previously clear separation of DMU_1 from all of the other DMUs is not maintained.

Undertaking the same operations with the values $w^2 = (1, 0.333, 0.167)$ and $w^3 = (1, 1, 0)$ as recorded for DMUs 2 and 3 in Table 3.5, we get the results displayed in Table 3.6.

Table 3.6 Initial solutions

	DMU_1	DMU_2	DMU_3
DMU	$h_j(w^1)$	$h_j(w^2)$	$h_j(w^3)$
1	1.000	0.667	0.250
2	1.000	1.000	0.500
3	0.667	1.000	1.000
4	0.750	0.857	0.500
5	0.857	0.750	0.333
6	0.500	0.500	0.250

Table 3.7 Data adjusted for 5% increment and decrement

	Efficient			Not efficient		
DMU	1	2	3	4	5	6
Output	1.00	1.00	1.00	1.00	1.00	1.00
Input 1	4.20	2.10	1.05	1.90	2.85	3.80
Input 2	1.05	2.10	4.20	2.85	1.90	3.80

Table 3.8 Results from 5% increment and decrement

	DMU_1	DMU_2	DMU_3
DMU	$h_j(w^1)$	$h_j(w^2)$	$h_j(w^3)$
1	0.953	0.635	0.238
2	0.953	0.952	0.476
3	0.635	0.952	0.952
4	0.790	0.902	0.526
5	0.902	0.789	0.351
6	0.526	0.526	0.263

As can be seen, a similar elevation to a top-rank tie occurs for DMU_2 but this does not displace DMU_2 for its top rank position. Finally the zero value for the v_2^* associated with DMU_3 does not cause any trouble and the very low value recorded for $h_6(w^3)$ is not very different from the value recorded in this same position in Table 3.2.

Following Thompson et al. (1994), we now introduce a 5% increment to the input values for every one of the efficient DMUs and a 5% decrement to the input values for every one of the inefficient DMUs in Table 3.1. Also like Thompson et al. (1994), we do not vary the outputs. The results are recorded in Table 3.7.

Applying the same procedure as before, we arrive at the new values displayed in Tables 3.5–3.8. As was case for Table 3.3, none of the originally top-ranked DMUs are displaced after these 5% increment adjustments to the data are made. In fact, just as was reported in the discussion following Table 3.3, a 15% increment–decrement change is needed before a displacement occurs for any top-ranked DMU.

We do not pursue this topic in further detail. Instead, we bring this discussion to a close by noting that no further changes occur in top rank until 20% increments for the inputs of the efficient DMUs and 20% decrements for the inefficient DMUs are made. The results are as displayed in Table 3.9 where, as can be seen,

Table 3.9 Results from 20% increments and decrements

	DMU_1	DMU_2	DMU_3
DMU	$h_j(w^1)$	$h_j(w^2)$	$h_j(w^3)$
1	0.870	0.580	0.217
2	0.870	0.870	0.435
3	0.580	0.870	0.870
4	0.882	1.000	0.588
5	1.000	0.882	0.392
6	0.588	0.588	0.294

DMU_5 has displaced DMU_1 from its top rank and DMU_4 has displaced DMU_2. The displacements differ from those obtained by Thompson et al. (1994) (as described in the discussion following Table 3.3). However, approximately the same degree of robustness is maintained. Indeed, as was also true in the Thompson et al. (1994) analysis even the 20% increments and decrements have not displaced DMU_3 from its top rank – as can be seen in column 3 of Table 3.9.

3.2.5 *Envelopment Approach*

The line of work we now follow extends the work of Charnes et al. (1992) to identify allowable variations in every input and output for every DMU before a change in status occurs for the DMU_o being analyzed. In contrast to the treatments in Thompson et al. (1994), the focus shifts to the "envelopment" rather than the "multiplier model" portrayed in (3.4). This shift of focus helps to bypass concerns that might be noted which arise from the possibility of different degrees of sensitivity that may be associated with alternate optima.

An easy place to begin is with the following formulation as first given in Zhu (1996a) and Seiford and Zhu (1998a)

Input orientation	Output orientation
$\beta_k^* = \min \beta_k$ for each $k = 1, \ldots, m$	$\alpha_l^* = \max \alpha_l$ for each $l = 1, \ldots, s$
$\sum_{j=1, j\neq 0}^{n} x_{kj}\lambda_j \leq \beta_k x_{ko}$	$\sum_{j=1, j\neq 0}^{n} y_{rj}\lambda_j \geq \alpha_\ell y_{lo}$
$\sum_{j=1, j\neq 0}^{n} x_{ij}\lambda_j \leq x_{io}, \quad i \neq k$	$\sum_{j=1, j\neq 0}^{n} y_{rj}\lambda_j \geq y_{ro}, \quad r \neq \ell$
$\sum_{j=1, j\neq 0}^{n} y_{rj}\lambda_j \geq y_{ro}, \quad r \neq 1, \ldots, s$	$\sum_{j=1, j\neq 0}^{n} x_{ij}\lambda_j \leq x_{io}, \quad i \neq 1, \ldots, m$
$\beta_k, \lambda_j \geq 0$	$\alpha_\ell, \lambda_j \geq 0.$ (3.12)

In this and subsequent developments, it is assumed that DMU_o is at least "weakly efficient" as defined in Chap. 1 and, as in Charnes et al. (1992), this DMU is omitted from the sums to be formed on the left in these expressions. Then singling out an input k and an output ℓ we can determine the maximum proportional change that can be allowed before a change in status from efficient to inefficient will occur for DMU_o. In fact the values of β_k^* and α_l^* provide the indicated boundaries for this input and output. Moreover, continuing in this fashion one may determine the boundaries for all inputs and outputs for this DMU_o.

Seiford and Zhu (1998a) also extend (3.12) to the following modified version of the CCR model

$$\begin{aligned}
&\beta^* = \min \beta \\
&\text{subject to} \\
&\sum_{j=1, j\neq 0}^{n} x_{ij}\lambda_j \leq \beta x_{io}, \quad i \in I \\
&\sum_{j=1, j\neq 0}^{n} x_{ij}\lambda_j \leq x_{io}, \quad i \notin I \\
&\sum_{j=1, j\neq 0}^{n} y_{rj}\lambda_j \geq y_{ro}, \quad r = 1, \ldots, s \\
&\beta, \lambda_j \geq 0.
\end{aligned} \tag{3.13}$$

Here, the set $i \in I$ consists of inputs where sensitivity is to be examined and $i \notin I$ represents inputs where sensitivity is not of interest.

Seiford and Zhu next use this model to determine ranges of data variation when inputs are worsened for DMU_o in each of its x_{io} and improved for the x_{ij} of every DMU_j, $j = 1, \ldots, n$ in the set $i \in I$. We sketch the development by introducing the following formulation to determine the range of admissible variations:

$$\sum_{j=1, j\neq o}^{n} x_{ij}\lambda_j \leq \delta x_{io}, \quad i \in I, \quad \text{where} \quad 1 \leq \delta \leq \beta^*. \tag{3.14}$$

Now assume that we want to alter these data to new values $x_{io} \geq x_{io}$ and $x_{ij} \leq x_{ij}$ which will continue to satisfy these conditions. To examine this case we use

$$\sum_{j=1, j\neq o}^{n} \hat{x}_{ij}\lambda_j \leq \delta \hat{x}_{io}, \quad \text{where} \quad 1 \leq \delta \leq \beta^* \tag{3.15.1}$$

and

$$\delta\hat{x}_{io} = x_{io} + \delta x_{io} - x_{io} = x_{io} + (\delta - 1)x_{io}$$

$$\hat{x}_{ij} = \frac{x_{ij}}{\delta} = x_{ij} + \frac{x_{ij}}{\delta} - x_{ij} = x_{ij} - \left(\frac{\delta - 1}{\delta}\right)x_{ij} \tag{3.15.2}$$

for every $j = 1, \ldots, n$ in the set $i \in I$.

Thus, $(\delta - 1)$ represents the proportional *increase* to be allowed in each x_{io} and $(\delta - 1)/\delta$ represents the proportional *decrease* in each x_{ij}. As proved by Seiford and Zhu, the range of variation that can be allowed for δ without altering the efficient status of DMU_o is given in the following.

Theorem 3.1 (Seiford and Zhu 1998a). If $1 \leq \delta \leq \sqrt{\beta^*}$ then DMU_o will remain efficient. That is, any value of δ within this range of proportional variation for both the x_{io} and x_{ij} will not affect the efficient status of DMU_o.

Here, δ is a parameter with its values to be selected by the user. Theorem 3.1 asserts that no choice of δ within the indicated range to be reclassified as inefficient when the x_{io} and x_{ij} defined in (3.15.2) are substituted in (3.13), because the result will still give $\beta^* \geq 1$.

Seiford and Zhu supply a similar development for outputs and then join the two in the following model which permits simultaneous variations in inputs and outputs:

$$\begin{aligned}
&\gamma^* = \min \gamma \\
&\text{subject to} \\
&\sum_{j=1, j\neq 0}^{n} x_{ij}\lambda_j \leq (1+\gamma)x_{io}, \quad i \in I \\
&\sum_{j=1, j\neq 0}^{n} x_{ij}\lambda_j \leq x_{io}, \quad i \notin I \\
&\sum_{j=1, j\neq 0}^{n} y_{rj}\lambda_j \geq (1-\gamma)y_{ro}, \quad r \in S \\
&\sum_{j=1, j\neq 0}^{n} y_{rj}\lambda_j \geq y_{ro}, \quad r \notin S \\
&\lambda_j \geq 0, \quad \gamma \text{ unrestricted}
\end{aligned} \tag{3.16}$$

where $i \in I$ represents the input set for which data variations are to be considered and $r \in S$ represents the output set for which data variations are to be considered. Using δ to represent allowable input variations and τ to represent allowable output variations, Seiford and Zhu supply the following.

Theorem 3.2 (Seiford and Zhu 1998a). If $1 \leq \delta \leq \sqrt{1+\gamma^*}$ and $\sqrt{1-\gamma^*} \leq \tau \leq 1$ then DMU_o will remain efficient.

Table 3.10 Comparison with Thompson et al.

	DMU_1	DMU_2	DMU_3
(g_o, g)	(41, 29)	(12, 11)	(41, 29)
SCSC1	20	14	20
SCSC2	32	9	32

Source: Seiford and Zhu (1998a). Here $g_o = \delta - 1$ and $g = (\delta - 1)/\delta$. See (3.15.1) and (3.15.2)

As Seiford and Zhu note, for $I = \{1, \ldots, m\}$ and $S = \{1, \ldots, s\}$, (3.16) is the same as the CCR correspond used in Charnes et al. (1996). Seiford and Zhu have thus generalized the Charnes et al. (1992) results to allow simultaneous variations in all inputs and outputs for every DMU in the sets $i \in I$ and $r \in S$.

We now turn to Table 3.10, which Seiford and Zhu used to compare their approach with the Thompson et al. (1994) approach. To interpret this table we note that all results represent percentages in the allowed data variation by applying these two different approaches to the data of Table 3.1. The values in the rows labeled SCSC1 and SCSC2 are secured from two alternate optima which Thompson et al. (1994) report as satisfying the strong complementary slackness condition. The parenthesized values of g_o and g at the top of Table 3.2 are reported by Seiford and Zhu as having been obtained by applying (3.13) to these same data.

The results from Seiford and Zhu seem to be more robust than is the case for Thompson et al., at least for DMU_1 and DMU_3. This is not true for DMU_2, however, where a 14% worsening of its inputs and a 14% improvement in the inputs of the nonefficient DMUs is required under the Thompson et al. (1994) approach before DMU_2 will change from efficient to inefficient in its status. However, the Seiford and Zhu approach shows that DMU_2 will retain its efficient status until at least a 12% worsening of its two inputs occurs along with an 11% improvement in these same inputs for the inefficient DMUs. A range of $12\% + 11\% = 23\%$ does not seem to be far out of line with the $2 \times 14\% = 28\%$ or the $2 \times 9\% = 18\%$ reported by Thompson et al. (1994). Moreover, as Seiford and Zhu note, their test is more severe. They match their worsening of DMU_2's inputs with improvement of the inputs of *all* of the other DMUs – including the efficient DMU_1 and $DMU_{2\_}$ – whereas Thompson et al. (1994) worsen the inputs of *all* of the efficient DMUs and improve the inputs of only the inefficient DMUs. (Note the fact that Seiford and Zhu deal only with "weak efficiency" is not pertinent here because DMU_1, DMU_2, and DMU_3 are all strongly efficient.)

Seiford and Zhu (1998b) also discuss situations where absolute (rather than proportional) changes in the data are of interest. This makes it possible to study sensitivity with the additive model discussed in Chap. 2 of this handbook. We here briefly discuss this approach. The absolute data variation can be expressed as

For DMU_o

$$\begin{cases} \hat{x}_{io} = x_{io} + \alpha_i \quad \alpha_i \geq 0, \quad i \in \mathbf{I} \\ \hat{x}_{io} = x_{io} \quad i \notin \mathbf{I} \end{cases} \quad \text{and} \quad \begin{cases} \hat{y}_{ro} = y_{ro} - \beta_r \quad \beta_r \geq 0, \quad r \in \mathbf{O} \\ \hat{y}_{ro} = y_{ro} \quad r \notin \mathbf{O} \end{cases}$$

For DMU_j $(j \neq o)$

$$\begin{cases} \hat{x}_{ij} = x_{ij} - \tilde{\alpha}_i \quad \tilde{\alpha}_i \geq 0, \quad i \in \mathbf{I} \\ \hat{x}_{ij} = x_{ij} \quad i \notin \mathbf{I} \end{cases} \quad \text{and} \quad \begin{cases} \hat{y}_{rj} = y_{rj} + \tilde{\beta}_r \quad \tilde{\beta}_r \geq 0, \quad r \in \mathbf{O} \\ \hat{y}_{rj} = y_{rj} \quad r \notin \mathbf{O} \end{cases}$$

where (^) represents adjusted data. Note that the data changes defined above are not only applied to all DMUs, but also different in various inputs and outputs.

Based upon the above data variations, Seiford and Zhu (1998b) provide the following model

$$\begin{aligned} & u^* = \min u \\ & \text{subject to} \\ & \sum_{j=1, j\neq 0}^{n} x_{ij}\lambda_j \leq x_{io} + u, \quad i \in I \\ & \sum_{j=1, j\neq 0}^{n} x_{ij}\lambda_j \leq x_{io}, \quad i \notin I \\ & \sum_{j=1, j\neq 0}^{n} y_{rj}\lambda_j \geq y_{ro} - u, \quad r \in \mathrm{s} \\ & \sum_{j=1, j\neq 0}^{n} y_{rj}\lambda_j \geq y_{ro}, \quad r \notin \mathrm{s} \\ & \sum_{j=1, j\neq 0}^{n} \lambda_j = 1 \\ & u, \lambda_j \geq 0, \quad \forall j. \end{aligned} \tag{3.17}$$

If $\mathbf{I} = \{1, 2, \ldots, m\}$ and $\mathbf{O} = \{1, 2, \ldots, s\}$, then model (3.17) is used by Charnes et al. (1992) to study the sensitivity of efficiency classifications in the additive model via the L_∞ norm when variations in the data are only applied to DMU_o.

Theorem 3.3 (Seiford and Zhu). Suppose DMU_o is a frontier point. If $0 \leq \alpha_i + \tilde{\alpha}_i \leq u^* \quad (i \in \mathbf{I})$, $0 \leq \beta_r + \tilde{\beta}_r \leq \gamma^* \quad (r \in \mathbf{O})$, then DMU_o remains as a frontier point, where u^* is the optimal value to (3.17).

If we change the objective function of (3.17) to "minimize $\sum_{i\in \mathbf{I}} r_i^- + \sum_{r\in \mathbf{O}} \gamma_r^+$," we obtain the following model which studies the sensitivity of additive DEA models discussed in Chap. 2:

$$
\begin{aligned}
&\min \sum_{i\in I} \gamma_i^- + \sum_{r\in O} \gamma_r^+ \\
&\text{subject to} \\
&\sum_{\substack{j=1\\ j\neq o}}^{n} \lambda_j x_{ij} \leq x_{io} + \gamma_i^- \qquad i \in \boldsymbol{I} \\
&\sum_{\substack{j=1\\ j\neq o}}^{n} \lambda_j x_{ij} \leq x_{io} \qquad i \notin \boldsymbol{I} \\
&\sum_{\substack{j=1\\ j\neq o}}^{n} \lambda_j y_{rj} \leq y_{ro} + \gamma_i^+ \qquad r \in \boldsymbol{O} \\
&\sum_{\substack{j=1\\ j\neq o}}^{n} \lambda_j y_{rj} \leq y_{ro} \qquad r \notin \boldsymbol{O} \\
&\sum_{\substack{j=1\\ j\neq o}}^{n} \lambda_j = 1 \\
&\gamma_i^-, \gamma_i^+, \lambda_j (j \neq o) \geq 0.
\end{aligned}
\tag{3.18}
$$

Based upon model (3.18), we have

Theorem 3.4 (Seiford and Zhu 1998b). Suppose DMU_o is a frontier point. If $0 \leq \alpha_i + \tilde{\alpha}_i \leq \gamma_i^{-*} \quad (i \in \mathbf{I}), 0 \leq \beta_r + \tilde{\beta}_r \leq \gamma_r^{+*} \quad (r \in \mathbf{O})$, then DMU_o remains as a frontier point, where $\gamma_i^{-*} \quad (i \in \mathbf{I})$ and $\gamma_r^{+*} \quad (r \in \mathbf{O})$ are optimal values in (3.18).

There are many more developments in these two approaches which we also do not cover here. We do need to note, however, that Seiford and Zhu (1998b, c) extend their results to deal with the infeasibility that can occur in such models as (3.12). Seiford and Zhu note that the possibility of infeasibility is not confined to the case when convexity is imposed. It can also occur when certain patterns of zeros are present in the data. They show that infeasibility means that the DMU_o being tested will preserve its efficient status in the presence of infinite increases in its inputs and infinite decreases in its outputs.

Although the above discussion is focused on changes in a specific (weakly) efficient DMU, Seiford and Zhu (1998b) and Zhu (2001) have extended the approach to situations where (unequal) changes in all DMUs (both efficient and inefficient DMUs) are considered. Further, Zhu (2009) provides a software that performs this type of sensitivity analysis.

3.3 Summary and Conclusion

Using data from a study evaluating performances of Chinese cities (Charnes et al. 1989) and a Chinese textile company (Zhu 1996b), Seiford and Zhu (1998a, b) conclude that their results show DEA results to be robust. This is the same conclusion that Thompson et al. (1994) arrive at from their sensitivity analysis of independent oil companies, Kansas farms, and Illinois coal mines.

This brings us back to the opening discussion in this chapter which distinguished "substantive findings" and "methodological approaches." These findings by Seiford and Zhu and Thompson et al. (1994) differ from substantive characterizations like those in Sexton et al. (1986) which we cited earlier. The latter believe that results in DEA would generally fail to be robust because of its reliance on extreme value observations, but do not seem to have tested this claim with actual data. Seiford and Zhu and Thompson et al. (1994) do at least supply the evidence (as well as rationales) for their conclusions. The evidence they adduce comes from only limited bodies of data, and, in addition, these analyses are all limited by their focus on only changes in the efficient status of different DMUs. In any case, we now have a collection of methods which can be used to determine the robustness of results in any use of DEA.

As we have observed, the progress in the sensitivity analysis studies we discussed has effected improvements in two important directions. First, this work has moved from evaluating one input or one output at a time in one DMU and has proceeded into more general situations where all inputs and outputs for all DMUs can be simultaneously varied. Second, the need for special algorithms and procedures has been reduced or eliminated.

Finally, Zhu (2009) provide a detailed discussion on "envelopment approach" with Microsoft® Excel-based software. This greatly enhances the applicability of the DEA sensitivity analysis approaches. Further, Cooper et al. (2000) point out that sensitivity analyses are not restricted to the approaches discussed in this chapter. The window analysis described in Chap. 1 of this handbook can also be treated as a method for studying the stability of DEA results because such window analyses involve the removal of entire sets of observations and their replacement by other (previously not considered) observations.

References

Ahn T, Seiford LM. Sensitivity of DEA to models and variable sets in an hypothesis test setting: The efficiency of university operations. In: Ijiri Y, editor. Creative and innovative approaches to the science of management. New York: Quorum Books; 1993.

Banker RD. Maximum likelihood, consistency and data envelopment analysis: a statistical foundation, this volume.

Banker RD, Chang H, Cooper WW. Simulation studies of efficiency, returns to scale and misspecification with nonlinear functions in DEA. Ann Oper Res. 1996;66:233–53.

Charnes A, Cooper WW. Management models and industrial applications of linear programmer. New York: John Wiley and Sons; 1961.
Charnes A, Cooper WW. Structural sensitivity analysis in linear programming and an exact product form left inverse. Naval Res Log Quart. 1968;15:517–22.
Charnes A, Cooper WW, Li S. Using DEA to evaluate relative efficiencies in the economic performance of Chinese cities. Socio-Econ Plann Sci. 1989;23:325–44.
Charnes A, Neralic L. Sensitivity analysis of the proportionate change of inputs (or outputs) in data envelopment analysis. Glasnik Matematicki. 1992;27:393–405.
Charnes A, Cooper WW, Lewin AY, Morey RC, Rousseau JJ. Sensitivity and stability analysis in DEA. Ann Oper Res. 1985;2:139–50.
Charnes A, Haag S, Jaska P, Semple J. Sensitivity of efficiency calculations in the additive model of data envelopment analysis. J Syst Sci. 1992;23:789–98.
Charnes A, Rousseau JJ, Semple JH. Sensitivity and stability of efficiency classifications in DEA. J Prod Anal. 1996;7:5–18.
Charnes A, Cooper WW, Thrall RM. A Structure for characterizing and classifying efficiencies in DEA. J Prod Anal. 1991;3:197–237.
Cooper WW, Seiford LM, Tone K. Data envelopment analysis: a comprehensive text with models references and DEA-solver software applications. Boston: Kluwer Academic Publishers; 2000.
Neralic L. Sensitivity in data envelopment analysis for arbitrary perturbations of data. Glasnik Matematicki. 1997;32:315–35.
Seiford LM, Zhu J. Stability regions for maintaining efficiency in data envelopment analysis. Eur J Oper Res. 1998a;108:127–39.
Seiford LM, Zhu J. Sensitivity analysis of DEA models for simultaneous changes in all of the data. J Oper Res Soc. 1998b;49:1060–71.
Seiford LM, Zhu J. Infeasibility of super-efficiency data envelopment analysis models. INFOR. 1998c;37(2):174–87.
Sexton TR, Silkman RH, Hogan RH. Measuring efficiency: an assessment of data envelopment analysis, new directions for program evaluations. In: Silkman RH, editor. Measuring efficiency: an assessment of data envelopment analysis. Publication No. 32 in the series New Directions for Program Evaluations. A publication of the American Evaluation Association. San Francisco: Jossy Bass; 1986.
Simar L, Wilson P. DEA Bootstrapping, this volume.
Thompson RG, Dharmapala PS, Diaz J, Gonzalez-Lima MD, Thrall RM. DEA multiplier analytic center sensitivity analysis with an illustrative application to independent oil Cos. Ann Oper Res. 1996;66:163–80.
Thompson RG, Dharmapala PS, Thrall RM. Sensitivity analysis of efficiency measures with applications to Kansas farming and Illinois coal mining. In: Charnes A, Cooper WW, Lewin AY, Seiford LM, editors. Data envelopment analysis: theory, methodology and applications. Massachusetts: Norwell Kluwer Academic Publishers; 1994. p. 393–422.
Wilson PW. Detecting influential observations in data envelopment analysis. J Prod Anal. 1995;6:27–46.
Zhu J. Robustness of the efficient DMUs in data envelopment analysis. Eur J Oper Res. 1996a;90:451–60.
Zhu J. DEA/AR analysis of the 1988–1989 performance of Nanjing textile corporation. Ann Oper Res. 1996b;66:311–35.
Zhu J. Super-efficiency and DEA sensitivity analysis. Eur J Oper Res. 2001;129(2):443–55.
Zhu J. Quantitative models for performance evaluation and benchmarking: data envelopment analysis with spreadsheets. Boston: Springer Science; 2009.

Chapter 4
Choices and Uses of DEA Weights

William W. Cooper, José L. Ruiz, and Inmaculada Sirvent

Abstract We review the literature of extensions and enhancements of the DEA basic methodology from the perspective of the problems that can be addressed by dealing with the dual multiplier formulation of the DEA models. We describe different approaches that allow incorporating into the analysis price information, reflecting meaningful trade-offs, incorporating value information and managerial goals, making a choice among alternate optima for the weights, avoiding non-zero weights, avoiding large differences in the values of multipliers, improving discrimination and ranking units. We confine attention to the methodological aspects of these approaches and show in many instances how others have used these approaches in applications in practise.

Keywords Data envelopment analysis • Weights

4.1 Introduction

Data Envelopment Analysis, as introduced in Charnes et al. (1978), is a methodology for the assessment of relative efficiency of a set of decision-making units (DMUs) that use several inputs to produce several outputs. In this seminal paper, the authors propose to solve, for each DMU_0, the following DEA model

J.L. Ruiz (✉)
Centro de Investigación Operativa, Universidad Miguel Hernández,
Avd. de la Universidad, s/n 03202-Elche, Alicante, Spain
e-mail: jlruiz@umh.es

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_4,

$$\begin{aligned} \text{Max} \quad & \frac{\sum_{r=1}^{s} u_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0}}, \\ \text{s.t.}: \quad & \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1, \quad j = 1, ..., n, \\ & v_i,\ u_r \geq 0, \qquad \forall i, r, \end{aligned} \tag{4.1}$$

which provides an efficiency score for the unit under assessment in the form of a ratio of a weighted sum of the outputs to a weighted sum of the inputs. By using the results on linear fractional problems in Charnes and Cooper (1962), model (4.1) can be converted into the following pair of dual linear problems

$$\begin{aligned} \text{Max} \quad & \sum_{r=1}^{s} u_r y_0, \\ \text{s.t.}: \quad & \sum_{i=1}^{m} v_i x_{i0} = 1, \\ & -\sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} u_r y_{rj} \leq 0, \quad j = 1, ..., n, \\ & v_i,\ u_r \geq 0, \qquad \forall i, r, \end{aligned} \tag{4.2}$$

which is called the dual multiplier formulation, and

$$\begin{aligned} \text{Min} \quad & \theta_0, \\ \text{s.t.}: \quad & \sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta_0 x_{i0}, \quad i = 1, ..., m, \\ & \sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{r0} \qquad r = 1, ..., s, \\ & \lambda_j \geq 0 \qquad \forall j, \end{aligned} \tag{4.3}$$

which is the primal envelopment formulation. These formulations are extended in Banker et al. (1984), with the so-called BCC model, which relaxes the assumption of constant returns to scale for the technology.

The term "multiplier" is used for the variables v_i and u_r in (4.2) in order to distinguish them from ordinary "weights" which are assumed a priori, whereas the variables in (4.2) are determined directly from the data. However, we shall use the terms interchangeably in our discussion.

With model (4.1), DEA provides a generalization of the so-called "engineering ratio," which is one of the most popular measures used to assess efficiency, but omits the assumption that we need to know the weights to be assigned on an a priori basis to each input and output for each DMU. These values are determined directly from the data by employing (4.2) where v_i is the weight assigned to input

$i, i = 1, ..., m,$ and u_r is the weight to be assigned to output $r, r = 1, ..., s,$ for $\text{DMU}_j, j = 1, ..., n$. (see Bulla et al. 2000 for further discussion of engineering uses of DEA). This DEA total weight flexibility is one of the most appealing aspects of this methodology. However, such flexibility often leads to unreasonable results in the sense that the weights provided are frequently inconsistent with the prior knowledge or accepted views on the involved production process. Another weakness of DEA has been pointed out in the literature in that large differences in the weights that a given DMU attaches to the different inputs and outputs may also be a concern. In an extreme situation, we have the case of zero weights: many DMUs take advantage of the weight flexibility and assess their efficiency by putting the weight on only a few inputs and outputs and ignore the remaining variables by assigning them a zero weight. It has also been claimed that it may sometimes be unaccepted that the different DMUs attach widely different weights to a given variable.

There is a body of research dealing with the problem of unacceptable weighting schemes. In this chapter, we review the literature of extensions and enhancements of the DEA methodology from the perspective of many of the problems that can be addressed with the dual multiplier formulation of the DEA models represented in (4.2). We focus on problems that can be dealt with the use of DEA weights and show how we can make specific choices of weights in order to provide suitable solutions for these problems. Obviously, some of these problems can also be dealt with other approaches that do not involve the use of DEA weights. When that is the case, this will be explicitly mentioned.

The chapter is organized into two parts, in which we distinguish between the methods that require some type of a priori information and those that need not use external information in order to approach such problems. It is to be noted that some of these methods can be used to approach more than one of the problems we discuss in different sections of the chapter. For example, using weight restrictions to incorporate value judgements regarding the relative importance of different variables may help improve discrimination and/or reduce weights dispersion. In each section, we emphasize the aspects of the approach analyzed that are of special interest to the problem addressed.

Most of the approaches we review in this chapter are based on the use of weight restrictions involving both individual inputs and/or outputs and pairs of variables, in terms of both absolute and virtual weights, which are to be incorporated into the multiplier formulation of the DEA model. In the case of inputs, for example the restrictions on individual weights can take the form $L_i \leq v_i \leq U_i$ or $L_i \leq v_i x_{i0} \leq U_i,$ and for pairs of variables we can use AR (assurance region) constraints (Thompson et al. 1986), of type I, $L_{ii'} \leq (v_i/v_{i'}) \leq U_{ii'},$ or type II, $L_{ir} \leq (v_i/\mu_r) \leq U_{ir},$ contingent weight restrictions (Pedraja-Chaparro et al. 1997), for example $L_{ii'} \leq (v_i x_{i0}/v_{i'} x_{i'0}) \leq U_{ii'},$ or constraints such as $L_i \leq v_i x_{i0}/\sum_{i=1}^{m} v_i x_{i0} \leq U_i$ (Wong and Beasley 1990). The setting of the bounds to be located in these constraints becomes one of the key issues in these approaches. Some other approaches use more general constraints, such as those in the cone-ratio models (Charnes et al. 1990), where $v = (v_1, ..., v_m)^{\mathrm{T}}$ and

$u = (u_1, ..., u_s)^{\mathrm{T}}$ are restricted to be in some specific, either separate or linked, polyhedral convex cones. Nevertheless, we are not concerned here with the theoretical aspects of the approaches used and the resulting models, their properties, interpretations and possible problems and weaknesses (these have been widely reviewed in other surveys; see, e.g. Allen et al. 1997; Thanassoulis et al. 2004). On the contrary, we confine attention to the methodological aspects behind the specification of the weight restrictions and their uses, which we do in many instances by showing how others have done it in applications reported in the literature.

Part I Using Prior Information and Expert Opinion

4.2 Using Price Information

When pricing information is known with precision, the measurement of efficiency may be defined in terms of these prices in order to provide measures of allocative and overall efficiency. However, in many cases, precise market price information is not available and only roughly related information is known, such as bounds on prices or certain relationships between them. In these cases, analysts have incorporated available information into DEA models to move from technical to overall efficiency. In the literature, this has been mainly done in the form of weight restrictions, the Input Cone Assurance Regions being the method mostly used.

Setting bounds for multiplier ratios in AR constraints when using price information has been done in different ways. A general approach mostly used is that based on determining a lower and an upper bound for the prices of each variable, $p_i^{\min}$ and $p_i^{\max}$, and then restricting the multiplier ratios according to these bounds as follows:

$$\frac{p_i^{\min}}{p_{i'}^{\max}} \leq \frac{\omega_i}{\omega_{i'}} \leq \frac{p_i^{\max}}{p_{i'}^{\min}}, \tag{4.4}$$

where ω_i represents the weight of the variable i, $i = 1, \ldots, m + s$, which can be either an input or an output.

The key in this approach is on determining ranges for the prices of each variable. It should be noted that this is not always an easy task, and historical data, statistics, relationships between variables and expert opinion are usually combined in order to derive them. Moreover, the analyst has to decide how to estimate $p_i^{\min}$ and $p_i^{\max}$ when several choices are meaningful [see, e.g. the discussion in Joro and Viitala (2004: 816), for a specific estimation of the production costs of the outputs among different alternatives]. In the simplest cases, $p_i^{\min}$ and $p_i^{\max}$ are set as either the minimum and maximum observed prices or $p_i^{\min} = (1 - p)\bar{p}_i$ and $p_i^{\max} = (1 + p)\bar{p}_i$, where $\bar{p}_i$ represents the average or median price and p is a

proportion that determines an allowable variation around $\bar{p}_i$ and is usually specified by the managers or other experts. A special case is that of the variables that measure either costs or earnings. In those situations, both $p_i^{\min}$ and $p_i^{\max}$ are usually set to be 1, presuming that each monetary unit paid or earned is worth this value.

The general approach (4.4) has been implemented with AR input multiplier constrains: see the efficiency analyses of Illinois coal mines in Thompson et al. (1995), the US's largest banks in assets size in Thompson et al. (1997), the commercial branches of a large Canadian bank in Paradi and Schaffnit (2004) performed with the model referred to as production model and that of the branches from a Portuguese commercial bank carried out by Camanho and Dyson (2005). This latter analysis was used to illustrate their method for the estimation of lower and upper bounds for the cost efficiency in situations of price uncertainty. In this analysis, the estimated bounds for the prices of each input had the particularity of being DMU-specific so that different constraints $p_{i0}^{\min}/p_{i'0}^{\max} \leq \omega_i/\omega_{i'} \leq p_{i0}^{\max}/p_{i'0}^{\min}$ were used depending on the DMU_0 evaluated.

Following (4.4), Joro and Viitala (2004) specify AR output multiplier constraints by combining expert opinion together with partial cost information in the evaluation of the efficiency of regional Forestry Boards in Finland.

Constraints such as (4.4) have also been used in the form of two separate input and output AR cones. See the efficiency analyses in Thompson et al. (1996) and Ray et al. (1998) dealing with major oil companies in USA and local major iron and steel Chinese firms, respectively. In the latter paper, the dual price system created by the Chinese economic reforms was used to determine $p_i^{\min}$ and $p_i^{\max}$ for all the variables but the labour force, where the minimum and maximum wage rates as derived from the total salary and worker benefits divided by the total number of workers by firm were used to set the range for the price of this input. The basic approach in (4.4) was slightly modified in Thompson et al. (1992) in order to specify two separate input and output AR cones in an efficiency analysis of oil/gas independent firms for each of the 7 years 1980–1986. In this study, two outputs [total production of crude oil (y_1) and natural gas (y_2)], and four inputs [total production cost (x_1), proven reserves of crude oil (x_2) and natural gas (x_3), and number of wells drilled (x_4)] were considered, and a different analysis was performed for each of the 7 years. The bounds for the only output multiplier ratio were not directly computed as in (4.4) but in a way detailed next that led to more accurate bounds and illustrates the situation in practise in which different choices of the limits for the multiplier ratios are available. To be specific, the authors divided the monthly minimum and maximum spot wellhead natural gas prices by the monthly average wellhead West Texas Intermediate crude oil price for each year, and then, the annual minimum and maximum gas/oil price ratios were estimated by averaging the corresponding monthly ratios. As they explain, if the respective minimum and maximum of the monthly gas/oil price ratios were taken, it would have not necessarily given minimum and maximum gas/oil price ratios in the same month. As for the AR input multiplier constrains, these were certainly specified as in (4.4) by using v_1 as the numeraire. Given that the first input

measures total cost, the value for this multiplier was specified to be 1, that is, $p_1^{\min} = p_1^{\max} = 1$; the reported minimum and maximum oil-equivalent values of the reserve additions by firm were used to set $p_2^{\min}$ and $p_2^{\max}$, respectively; these latter quantities were multiplied times the previously obtained bounds for the output multiplier ratio relating the prices of gas and oil to estimate $p_3^{\min}$ and $p_3^{\max}$; and $p_4^{\min}$ and $p_4^{\max}$ were specified as the onshore and offshore drilling cost data.

And finally, the general approach (4.4) has also been used to specify linked-cone AR (LC-AR) in profit ratio analyses. This is the case of the previously mentioned papers by Thompson et al. (1995, 1996, 1997), where the same values $p_i^{\min}$ and $p_i^{\max}$ used to derive the previously mentioned AR-I constraints were also used in (4.4) to relate the input and the output multipliers in form of AR-II constraints.

We would also like to stress that in some situations the analysts do not have information enough to determine precise lower and upper bounds for the ratios of multipliers. Nevertheless, in such situations, general knowledge, either practical or theoretical, on economic prices can at least be used to set the prices of certain variables so that they are not lower than those of other variables. As Kuosmanen and Post (2001) state, "one source of price information is prior knowledge on the quality or risk of the different inputs and outputs." For example, in the empirical application of their method for deriving lower and upper bounds for Farrell's cost overall efficiency in FDH technologies to commercial banks, these authors incorporate the constraint $v_1 - v_2 \geq 0$ concerning the relative price of equity capital (x_1) and debt capital (x_2), respectively, as the result of considering the assumption of economic theory that states that the cost of equity capital exceeds that of debt capital. In Joro and Viitala (2004), the authors restrict the output weights to the same weak ordering as the estimates of the production costs of the outputs to illustrate an alternative way to (4.4) to incorporate price information in the efficiency model.

4.3 Reflecting Meaningful Trade-Offs

In DEA, the production possibility set (PPS) is estimated from the data of all the DMUs together with some general axioms assumed for the underlying technology (see Banker et al. 1984). However, if information on realistic technological trade-offs is available, then it can be used to enlarge the set of possible input–output combinations beyond the traditional estimation of the PPS in order to obtain sharper technical efficiency measures based on more realistic frontiers.

When determining the trade-off information to be incorporated into the DEA model used, caution must be taken to ensure that the trade-offs actually reflect simultaneous changes in the levels of the corresponding variables without affecting the levels of the remaining inputs and outputs that are valid at all units in

the technology, since this information is generally used to estimate the whole PPS. As Podinovski (2004) states, the relation described by the used trade-offs should be relatively undemanding and, because of this, applicable to different units. Thus, the notion of trade-offs used in this context is different from the concept of marginal rates of substitution used in production economics, which represents the exact proportions in which the inputs and outputs of a particular unit on the efficient frontier can change and are generally different for different units.

This trade-off information has been traditionally incorporated into the DEA models in the form of weight restrictions, both of type AR-I and AR-II (see Podinovski 2004 for an implementation by modifying the primal envelopment formulation). It should be noted that the inclusion of AR constraints based on either price or trade-off information has a different impact on the resulting efficiency score. While in the latter case, the efficiency score retains its meaning of technical efficiency measure given that only objective information about the process is considered, in the former no direct information on the production process but on the economic values of the variables is considered so that the resulting efficiency score evaluates allocative in addition to technical efficiency.

Schaffnit et al. (1997) incorporate trade-off information based on standard transaction and maintenance times in their analysis of the Ontario-based branches of a large Canadian bank. Two models were used in that paper to evaluate the staff performances. One model aimed at evaluating the global staff performance using five inputs measuring the number of people of five types of staff and nine outputs to evaluate the different services provided. These consisted of the number of transactions processed of six types and the number of accounts of three types (as a proxy of the three considered maintenance activities). The second model focused on the assessment of transaction efficiency and used the same five inputs and only the six outputs concerning transactions. Assuming that working time can be indistinctly devoted to any of the activities described in outputs, and taking into account that inputs considered actually measure the number of effective hours of work, the authors use the transactions and maintenance times to define possible trade-offs between outputs. To be specific, AR constraints for the output multipliers as in (4.4) were incorporated into both models, but in this case $p_i^{\min}$ and $p_i^{\max}$ represented the range for the time required by a given type of transaction or maintenance activity and were set from the managers' opinion as $p_i^{\min} = (1 - 0.25)\bar{p}_i$ and $p_i^{\max} = (1 + 0.25)\bar{p}_i$, $\bar{p}_i$ being the corresponding standard time of either transaction or maintenance activity measured by the bank. The authors referred to them as the refined models. An analysis of the sensitivity to the allowable variation around $\bar{p}_i$, initially set at $p = 0.25$, was performed.

Olesen and Petersen (2002) used the probabilistic assurance regions (Olesen and Petersen 1999) in a different approach to incorporate trade-off information into DEA models. They performed an efficiency analysis of Danish hospitals in which one input, observed cost, and 483 outputs that measure the number of discharges within 483 patient categories were considered. The authors argue that

the input is homogeneous for many subgroups of the services that constitute a treatment in each of these patient categories and can be reallocated to other activities based upon measures of the average resource consumption. Thus, for each pair of patient categories they use the ratio of the corresponding average resource consumption as a measure of a possible rate of substitution. Based on distributional characteristics of measures of the average resource consumption in each patient category, all the output multipliers were also constrained by ARs as in (4.4). To be specific, $p_i^{\min}$ and $p_i^{\max}$ were estimated here as the left and right end points in confidence intervals for costs of each patient category.

Podinovski (2004) proposed a specific approach to translate production trade-offs into equivalent weight restrictions which we briefly describe next. As said before, a trade-off reflects certain simultaneous changes in the levels of inputs and/or outputs that are possible without affecting the levels of the remaining inputs and outputs. Hence, a trade-off can be described by means of a pair of vectors (P, Q), where the vector $P \in \mathbb{R}^m$ reflects the changes in the inputs and $Q \in \mathbb{R}^s$ the changes of the outputs. For example, suppose that there are two inputs and three outputs and that the reduction of 1 U of the first input is deemed technologically possible if the second input is increased by 2 U (without affecting the levels of the remaining variables). This trade-off can be described as $P = (-1, 2)^{\mathrm{T}}$ and $Q = (0, 0, 0)^{\mathrm{T}}$. If an increase of 1 U of the second output is deemed technologically possible provided that the first input is increased by 3 U, then $P = (3, 0)^{\mathrm{T}}$ and $Q = (0, 1, 0)^{\mathrm{T}}$.

Suppose that we have specified K different trade-offs (P_t, Q_t), $t = 1, ...K$, then, Podinovski proposes to incorporate these trade-offs into the analysis by adding the following weight restrictions to the dual multiplier formulation of the DEA model:

$$u^{\mathrm{T}} Q_t - v^{\mathrm{T}} P_t \leq 0, \quad t = 1, 2, ..., K, \tag{4.5}$$

where $u = (u_1, \ldots, u_s)^{\mathrm{T}}$ and $v = (v_1, \ldots, v_m)^{\mathrm{T}}$. By using this procedure, the examples above would result in the following two constraints: $v_1 - 2v_2 \leq 0$ and $u_2 - 3v_1 \leq 0,$, respectively.

Khalili et al. (2010) have recently proposed a new model that deals with difficulties of the AR-II constraints regarding the possible under-estimation of relative efficiency and infeasibilities. They illustrate its use with an application to the assessment of secondary schools in Portugal in which trade-off information is incorporated into the analysis following the approach in Podinovski (2004) in order to convert it into weight restrictions. The model used to evaluate the efficiency has three inputs characterizing the student cohort on entry of the secondary education and two outputs reflecting academic achievement on exit. As for the specification of trade-offs, values for the variables characterizing an average school were presented to an expert, who was asked to provide values for these variables for an (unobserved) school that can be considered equivalent in terms of performance relative to the average school. The expert provided four schools which differed

from the average school, and the differences between the values of these four schools and those of the average school were used to specify four different trade-offs. These were converted into AR-II constraints (note that, in contrast to the two previous papers, which only considered simultaneous changes in the outputs, linked trade-offs between inputs and outputs were used here) and incorporated into the new non-linear model proposed.

4.4 Incorporating Value Information and Managerial Goals

The freedom of DEA models to select weights in real-world applications may lead to weighting schemes that are not consistent with prior knowledge or accepted views and beliefs or that are quite unrealistic from a managerial point of view. To deal with these difficulties, we usually proceed by incorporating information concerning the relative importance of the inputs and outputs involved or some managerial preferences into the basic DEA model, if these are available. The resulting efficiency scores then evaluate both the technical inefficiency that arises from not fully exploiting production possibilities and the inefficiency due either to lack of fulfilment of managerial goals or to the departure from the specified value system of the inputs and outputs.

Incorporating value information or preferences into an analysis has been mostly done by using weight restrictions in the dual multiplier formulation of the DEA models. The key issue in this approach is in setting the bounds to be located in the constraints so that they reflect the information obtained from expert opinion.

Sometimes the experts are able to quantify the value of the variables and this can be straightforwardly used to set bounds for the corresponding ratios of weights. This is the case, for example, in the strategic model in Paradi and Schaffnit (2004) to evaluate the performance of commercial branches in line with the organization's goals. The managers give a value p_r to each output y_r and constraints like (4.4) on the relative importance of the outputs were specified, the individual range for each multiplier being $p_r \pm 20\%$.

In most situations, eliciting weight restrictions concerning the relative value of the variables involves a somewhat more elaborate process. Let us take as an example the comparison of university departments in Beasley (1990), where different types of weight restrictions were used. We here describe only some of these types of constraints as representative cases and not the full set of weight restrictions in the model that was eventually used in the analysis. In this study expert opinions, personal beliefs and also some external information were handled with the aim of specifying these weight constraints. The analysis was based on three inputs (x_1 = general expenditure, x_2 = equipment expenditure and x_3 = research income) and eight outputs (y_1 = number of undergraduates, y_2 = number of postgraduates on taught courses, y_3 = number of postgraduates doing research, y_4 = research income, and y_5, y_6, y_7 and y_8 being variables taking the value 1 if the department is rated star, A+,

A, A−, respectively, or 0 otherwise). To incorporate the general agreement that the weight attached to a postgraduate doing research should be greater than or equal to the weight attached to a postgraduate on a taught course and correspondingly for undergraduates, some restrictions of type $u_1 \leq u_2 \leq u_3$ were used. Beasley actually went further by replacing these constraints with $1.25^2 u_1 \leq 1.25 u_2 \leq u_3$ and $u_3 \leq 2u_1$. With respect to equipment expenditure, he incorporated the belief that its associated weight should reflect the total amount spent on equipment for the entire set of departments and expressed as a fraction of total value, $F = \sum_{j=1}^{n} x_{2j} / \sum_{j=1}^{n} \left(x_{1j} + x_{2j} + x_{3j}\right)$. In order to implement this more realistically, a flexibility of $\pm 20\%$ around this value was allowed, which resulted in the following constraint $0.8 \times F \leq \sum_{j=1}^{n} v_2 x_{2j} / \sum_{j=1}^{n} \left(v_1 x_{1j} + v_2 x_{2j} + v_3 x_{3j}\right) \leq 1.2 \times F$. A different approach was used to reflect the relative importance attached to teaching and research. The analysts in that case used the information provided by the Croham review of University Grants Committee, which revealed that 63.75% of the grants made to universities is for teaching, with the remainder being for research. This was taken as indicative of the relative importance attached to teaching (student numbers) and research by policy makers, and was implemented by restricting the proportion of total output associated with student numbers by means of the following constraint $0.8 \times 0.6375 \leq \sum_{r=1}^{3} u_r y_{r0} / \sum_{r=1}^{8} u_r y_{r0} \leq 1.2 \times 0.6375$, not only for the department under assessment but also for the entire set of departments considered as a whole $0.8 \times 0.6375 \leq \sum_{r=1}^{3} u_r \sum_{j=1}^{n} y_{rj} \Big/ \sum_{r=1}^{8} u_r \sum_{j=1}^{n} y_{rj} \leq 1.2 \times 0.6375$ (note again that a flexibility of 20% around the percentage 63.75 was allowed). At this point, the authors suggested the possibility of incorporating this constraint for all the individual departments in order to ensure that all of them satisfied this condition.

It should be noted the use of both absolute weight restrictions and restrictions on virtual weights in the previous paper. Virtuals allow us to avoid problems with the units of measurement, and they also have a more clear interpretation in terms of the importance attached to a given variable by the unit under assessment. However, virtual weight restrictions are unit-specific. As a consequence, if one was interested in incorporating into the model, a constraint such as $L_r \leq u_r y_{r0} / \sum_{r=1}^{s} u_r y_{r0} \leq U_r$ not only for the unit under assessment but also for all the DMUs this would be computational expensive. In such situations, Wong and Beasley (1990) suggest adding constraints of the type $L_r \leq u_r \left(\sum_{j=1}^{n} y_{rj}/n\right) \Big/ \sum_{r=1}^{s} u_r \left(\sum_{j=1}^{n} y_{rj}/n\right) \leq U_r$ to the set of constraints for the unit under assessment, which involves the use of an average DMU. See Sarrico and Dyson (2004) for discussions.

The contingent weight restrictions (Pedraja-Chaparro et al. 1997) offer another possibility in the use of weight restrictions involving virtuals. Cooper et al. (2009) use them in order to incorporate the opinion of a team coach of the Spanish basketball league (the ACB league) regarding the relative importance of the factors considered in their assessment of effectiveness of basketball players. For example, in the case of playmakers, the expert believed that the importance attached to rebounds should not be greater than that of field goal, and consequently, the following constraints were added $\left(u_{\mathrm{REB}} y_{\mathrm{REB},j} / u_{\mathrm{FG}} y_{\mathrm{FG},j}\right) \leq 1, \quad j = 1, \ldots, 41,$

for the 41 playmakers in the sample. Eventually, this is equivalent to using the following AR-I constraint $u_{\mathrm{REB}}/u_{\mathrm{FG}} \leq \min\{y_{\mathrm{FG},j}/y_{\mathrm{REB},j}, \quad j = 1, ..., 41\}$. Similar weight restrictions were also used with other pairs of the variables that described the different aspects of the game (shooting, ball handling or defense).

A different approach to the specification of weight restrictions regarding the relative importance of the variables can be found in Sarrico et al. (1997), which reports a case study concerned with the university selection in the UK. They used factors that had been previously classified into three categories: very important factors (V), important factors (I) and less important factors (L). Each very important factor might be weighted heavier than each important one and those in turn heavier than each less important one but this was discarded since it would require a large number of weight restrictions in the cases of having several factors in each category. The authors were able to retain this general idea and avoid these difficulties by constructing ranges of virtual weights for each category of factors in the following manner: The less important factors were not allowed to have virtual weights bigger than α: $u_{\mathrm{L}}y_{\mathrm{L0}} \leq \alpha$; factors considered important would be allowed to have weights in the range α to 2α: $\alpha \leq u_{\mathrm{I}}y_{\mathrm{I0}} \leq 2\alpha$ and the very important ones allowed weights of at least 2α: $u_{\mathrm{V}}y_{\mathrm{V0}} \geq 2\alpha$. α was calculated by imposing that the aggregate of the lower bounds on virtuals determined by 50% of the aggregate performance measure while retaining 50% for weight flexibility: $i \times \alpha + v \times 2\alpha = 0.5$, where i and v are the number of important and very important factors, respectively. See also Sarrico and Dyson (2000) for a similar approach.

In some cases, the bounds of weights restrictions are expressed by using a parameter and the sensitivity of the results to its specification is to be subsequently analyzed. For example, Shimshak et al. (2009) claim that it is essential to include measures of quality care in the evaluation of nursing homes. However, this is often ignored (as has happened in other applications reported in the literature). Instead they use the following weight restrictions $\min_i \left(u_i^{\text{quality}}\right) \geq \theta \times \max_j \left(u_j^{\text{quantity}}\right)$, where θ is a parameter to be chosen by management reflecting their assessment of the relative importance of quantity and quality aspects of performance. The results with different choices of θ are analyzed. In the measurement of performance of nations at the Summer Olympics in Lozano et al. (2002), the authors also use a couple of parameters α, β in two AR-I constraints, $(u_{\mathrm{NG}}/u_{\mathrm{NS}}) \geq \alpha$ and $(u_{\mathrm{NS}}/u_{\mathrm{NB}}) \geq \beta$, where u_{NG}, u_{NS} and u_{NB} are the weights attached to the number of gold medals, silver medals and bronze medals, respectively, in order to reflect how much worthier a gold medal is than a silver medal and this latter than a bronze medal. The sensitivity of the results to the choice of these parameters is also analyzed.

A more sophisticated technique such as the analytic hierarchy process (AHP) (Saaty 1980) has also been used in order to indirectly specify value judgements from experts regarding the relative value of the variables involved in the efficiency analysis, which is further used to derive weight restrictions to be added to the DEA models. In a first step, the experts make pairwise comparisons

between each couple of inputs (outputs) and their judgement is quantified and recorded in a matrix. AHP transforms the information in this matrix into a weight vector $\omega^k = \left(\omega_1^k, \ldots, \omega_m^k\right)$, $\sum_{i=1}^m \omega_i^k = 1$, which represents the relative importance of each input in the efficiency evaluation (analogously for the outputs). The information provided by these weights has been used for setting bounds on weight restrictions in different manners. In Takamura and Tone (2003), Lee et al. (2009) and Ramón et al. (2010c), the authors use constraints in the form $L_{ij} \leq (v_i/v_j) \leq U_{ij}$, where $L_{ij} = \min_k (\omega_i^k/\omega_j^k)$ and $U_{ij} = \max_k (\omega_i^k/\omega_j^k)$, k being the superscript associated with the experts involved. And in Shang and Sueyoshi (1995), the restrictions used had the form $L_i \leq (v_i x_{i0}/\sum_{i=1}^m v_i x_{i0}) \leq U_i$, where $L_i = \min_k \omega_i^k$ and $U_i = \max_k \omega_i^k$, ω_i^k representing in that case the importance of input i in relation to the improvement of output k in the opinion of the unique expert involved in this analysis.

It should be noted that the expert opinions do not always allow setting bounds on weight restrictions. Nevertheless, it may occasionally be processed in a different manner that makes it possible to take value judgements into consideration. In the analysis in Brockett et al. (1997), the regulatory officials were not able to set bounds on specific variables, which made it impossible to use an AR (or like) approaches. However, their expertise was utilized to identify a collection of "excellent" banks (used as "model" DMUs) and a cone-ratio model was used with the multipliers of these "excellent" banks (in the unbounded model) as admissible directions. To be precise, the following constraints are to be added to (4.2): $v \in V$ and $u \in U$, where V and U are the polyhedral convex cones spanned by the vectors of optimal input and output weights chosen by the "excellent" banks, i.e. $V = A^{\mathrm{T}}\alpha$, where $A_{k\times m}$ is a matrix having as rows the optimal input weights of the k DMUs chosen by the experts and $\alpha_{k\times 1} \geq 0$, and $U = B^{\mathrm{T}}\beta$, where $B_{k\times s}$ is a matrix having as rows the optimal output weights of the k DMUs chosen by the experts and $\beta_{k\times 1} \geq 0$. Note that this approach ignores the existence of alternate optima for the weights.

Finally, before ending this part of the chapter, it is worth mentioning that in practise we sometimes have available both price information and value information simultaneously while on other occasions this information includes both prices and trade-offs. If that is the case, all the information can be incorporated into the DEA model through the corresponding weight restrictions. See the case of the strategic model in Paradi and Schaffnit (2004), which includes both input multiplier constraints based on salary ranges offered by management and output multiplier constraints based on management's value system, which yields to what the authors interpret as an overall cost-effectiveness measure. And also the models in Schaffnit et al. (1997) that look at the cost-minimizing behaviour of the branches, which result from adding price information of the inputs in the form of the constraints in (4.4) to the previously mentioned refined models that contain trade-off information.

Part II Using Information in Data

In this part of the chapter, we present some approaches to problems that can be dealt with methods that do not require any a priori information to be used and, thus, rely on the information provided by the data. Important sources of trouble in the use of the dual models of DEA are represented by: (1) the existence of alternate optima, which raises questions as to which multiplier values to use and (2) the presence of numerous zeros that hamper usages for substitution and transformation analyses and also mean that some inputs and/or outputs initially considered by the analyst have been ignored in the analysis. Aside from the zero weights, the differences in the values of the multipliers, both those within a given solution for a DMU and the differences between the weight profiles of the different DMUs, have been claimed in the literature as some issues that deserve to be investigated. We also address the problem of lack of discrimination, strongly related to that of ranking units, with methods based on some specific choices of weights. It should be noted that all these problems can also be addressed with some of the approaches we discussed in the first part of this chapter. For example, incorporating value judgements concerning the relative importance of some variables by means of, say, the inclusion of some AR constraints might lead to non-zero weights, reduce the variation in weights and improve the discrimination among the efficient DMUs. The methods reviewed here have been designed specifically to address these problems and they do not require any type of a priori information since they only use that contained in the data.

4.5 Choosing From Alternate Optima

Efficiency analyses are mainly concerned with obtaining the efficiency scores and the sources and amounts of any inefficiencies of each member of the set of DMUs to be assessed and also providing the corresponding peers which can be used as benchmarks. However, analysts are sometimes interested in additionally estimating the weights with the purpose of obtaining added insight into the relative value of the inputs and outputs involved in such analyses. A problem that often arises in practise, especially in the case of the extreme efficient units, is that we may have different optimal weights associated with the efficiency score of a given DMU and this may provide very different insights into the role played by the variables used in the efficiency assessment. In practise, analysts generally use the weights provided by the software used, overlooking the possibility that there may be other optimal weights leading to very different conclusions. For this reason, in Cooper et al. (2007), the authors claim there is a need for selecting weights from among the optimal solutions of the dual multiplier models according to some criteria.

It should be noted that inefficient units in practise generally either are on the weak efficient frontier or are projected on to it [i.e. they belong to $F \cup NF$ according to the classification of DMUs in Charnes et al. (1991)] and have unique optimal weights in the DEA model, while the extreme efficient units (those in set E) always have infinite alternative optimal solutions for their weights. This is why the procedures for the selection of weights among alternate optima usually focus on the extreme efficient DMUs.

The selection of weights for a given extreme efficient DMU proposed in Cooper et al. (2007) is connected with the dimension of the efficient facets of the frontier and involves two steps. In the first step, this approach selects weights associated with the facets of higher dimension that this unit generates and, in particular, it selects those weights associated with a full dimensional efficient facet (FDEF) if any. In this sense, the weights provided by this procedure have the maximum support from the observed data since they are associated with hyperplanes that maximize contact with the PPS. These weights are the optimal solutions of the following model, which selects between the optimal weights of a given DMU_0 in E those associated with a hyperplane that is supported by the maximum possible number of extreme efficient units

$$
\begin{aligned}
\text{Min} \quad & I_0 = \sum_{j \in E} I_j, \\
\text{s.t.:} \quad & \sum_{i=1}^{m} v_i x_{i0} = 1, \\
& -\sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} u_r y_{rj} + t_j = 0, \quad j \in E, \\
& \sum_{r=1}^{s} u_r y_{r0} = 1, \\
& t_j - M I_j \leq 0, \qquad j \in E, \\
& I_j \in \{0, 1\}, \qquad j \in E, \\
& v_i,\ u_r,\ t_j \geq 0, \qquad \forall i, r, j,
\end{aligned}
\tag{4.6}
$$

where M is a big positive value, bigger than any positive real number.

In particular, DMU_0 is on a FDEF if the optimal value of (4.6) equals $|E| - (m + s - 1)$ (this model is similar to that proposed in Olesen and Petersen (1996) as a test for the existence of FDEFs). The second step of the procedure complements the selection of weights by choosing among the optimal solutions of (4.6) those weights that maximize the relative value of the involved inputs and outputs. To do this, we simply need to maximize z_0 subject to the constraints $v_i x_{i0} \geq z_0,\ i = 1, \ldots, m$, and $u_r y_{r0} \geq z_0, r = 1, \ldots, s$, where v_i and u_r represent the set of optimal solutions of (4.6). This approach was used in an assessment of effectiveness of basketball players in Cooper et al. (2009) which had the purpose of providing a more comprehensive portrayal of the performance

Table 4.1 Data of example (Wong and Beasley, 1990)

	Inputs			Outputs			
DMU	x_1	x_2	x_3	y_1	y_2	y_3	Efficiency
1	12	400	20	60	35	17	1
2	19	750	70	139	41	40	1
3	42	1,500	70	225	68	75	1
4	15	600	100	90	12	17	0.820
5	45	2,000	250	253	145	130	1
6	19	730	50	132	45	45	1
7	41	2,350	600	305	159	97	1

Table 4.2 Results of example

DMU		v_1	v_2	v_3	u_1	u_2	u_3	Dimension
1	CR&S	6.412	0.049	0.164	0.603	1.730	0.193	4 (1, 5, 6, 7)
	DEA (LINDO)	0	0	5	0	2.857	0	1 (1)
2	CR&S	4.758	0.006	0.069	0.554	0.445	0.120	3 (2, 6, 7)
	DEA (LINDO)	5.233	0	0.008	0.719	0	0	2 (2, 7)
3	CR&S	0.612	0.011	0.818	0.076	0.250	0.880	3 (1, 3, 6)
	DEA (LINDO)	0	0	1.429	0.444	0	0	1 (3)
5	CR&S	1.354	0.016	0.031	0.139	0.395	0.059	4 (1, 5, 6, 7)
	DEA (LINDO)	1.945	0	0.050	0	0.690	0	2 (5, 7)
6	CR&S	3.901	0.029	0.101	0.364	1.043	0.112	4 (1, 5, 6, 7)
	DEA (LINDO)	0	0,128	0.132	0	0	2.222	2 (5, 6)
7	CR&S	0.962	0.022	0.013	0.123	0.344	0.080	4 (1, 5, 6, 7)
	DEA (LINDO)	1.774	0	0.045	0	0.629	0	2 (5, 7)

of the effective players.To illustrate the use of the procedure described above, we use the data in Table 4.1 that have been taken from Wong and Beasley (1990) and consist of seven DMUs with three inputs and three outputs. An additional column showing the efficiency scores is included.

There are six extreme efficient units and one inefficient unit (DMU 4). As usual, the inefficient unit belongs to class NF and has unique optimal weights while the efficient ones belong to class E and have infinite optimal weights. The results of the procedure in Cooper et al. (2007) are shown in Table 4.2 (with the label CR&S), together with those provided by the classical CCR input-oriented model, which have been obtained by solving (4.2) with conventional software (LINDO), for comparisons. For each solution, we also record the dimension of the associated efficient facet and, in brackets, the units that span it. We can see that when no additional criterion is used to select weights, the support from observed data of the facets used to evaluate the efficiency is generally low. Only one or two efficient units contribute to span each of these facets. In contrast, CR&S, which maximizes the support, selects facets of the PPS with either three or four efficient units spanning them (note that no FDEF exists in this PPS). Moreover, the weight profiles provided by the classical DEA model

assign zero weights to at least half of the inputs and outputs for all the efficient units while CR&S always yields non-zero weights.

Trying to maximize the importance (or contribution) of the variables involved in the analysis is one of the criteria that has also been used in the literature on the choice of weights in different contexts. For example, Thanassoulis (2001) provides a model that also maximizes z_0 subject to the constraints $v_i x_{i0} \geq z_0$, $i = 1, \ldots, m$, and $u_r y_{r0} \geq z_0$, $r = 1, \ldots, s$, where v_i and u_r in that case simply represent optimal solutions of the CCR model for DMU_0. This model is directed to providing weights regarding what he calls "robustness" of the efficiency rating and is mainly applied to efficient units. The optimal value of this model allows us to judge the extent to which the assessed unit relies on a limited number of inputs and outputs to get its maximum efficiency rating. In the context of the choice of weights to be used in cross-efficiency evaluations, the "neutral" DEA model proposed in Wang and Chin (2010a) includes the constraints $u_r y_{r0} \geq z_0$, $r = 1, \ldots, s$, for the outputs, where z_0 is to be maximized, and so, we have weights in which outputs can be made as much use as possible.

Other criteria that could be used to the choice of weights among alternate optima would include those used as alternative secondary goals in cross-efficiency evaluations. See, for example the well-known benevolent and aggressive approaches (Sexton et al. 1986; Doyle and Green 1994a), which select between the optimal weights of a given DMU those that globally maximize and minimize (respectively) the efficiency scores of all the DMUs, while maintaining its self-efficiency rating. This would also include the model used in the first stage of the two-step procedure in the multiplier bound approach to the assessment of relative efficiency without slacks in Ramón et al. (2010a), which looks for the optimal solutions with the least dissimilar weights, and its use has also been extended to cross-efficiency evaluations in Ramón et al. (2010b) (more details of these approaches are provided in the next sections).

Finally, Charnes et al. (1991) propose a procedure that provides an optimal solution of the CCR model satisfying strong complementary slackness condition (SCSC) which can also be used to select weights for each efficient DMU. For a given DMU_0, this method starts by solving the following model:

$$
\begin{aligned}
\text{Max} \quad & \sum_{j=1,\, j\neq 0}^{n} t_j + \sum_{r=1}^{s} u_r + \sum_{i=1}^{m} v_i, \\
\text{s.t.:} \quad & \sum_{i=1}^{m} v_i x_{i0} = 1, \\
& \sum_{r=1}^{s} u_r y_{r0} = 1, \\
& -\sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} u_r y_{rj} + t_j = 0, \qquad \forall j, \\
& t_j, v_i, u_r \geq 0, \qquad \forall j, i, r. \qquad (4.7)
\end{aligned}
$$

Table 4.3 Weights from SCSC solutions by using model (4.7)

DMU	v_1	v_2	v_3	u_1	u_2	u_3
1	2.274	0.044	2.749	0.520	1.642	0.668
2	3.179	0.050	0.035	0.625	0.155	0.170
3	0.491	0.009	0.934	0.222	0.075	0.598
5	1.908	0.001	0.046	0.054	0.108	0.544
6	2.243	0.046	0.481	0.201	0.439	1.194
7	1.819	0.007	0.013	0.101	0.335	0.163

Then, the positive variables in the solution obtained are removed from the objective and the model is re-run. This process is repeated until no variable remains in the objective. The average of all the solutions (v_i, u_r) in each step is the SCSC solution proposed. It is to be noted that this procedure guarantees non-zero weights. Table 4.3 shows the results provided by this method by using again the data in Table 4.1. As can be seen, all the weights are strictly positive. We note that, as a result of averaging alternative weight profiles, the dimension of the efficient facets associated with these SCSC solutions is 1. That is, only the evaluated efficient DMU is located on the corresponding facet.

The existence of alternative optimal weights is also a problem when estimating marginal rates of substitution or transformation (MR) for the efficient units. To overcome this problem, some procedures based on an implicit selection of optimal weights have been proposed, which gives rise to the desired MRs. Rosen et al. (1998) provide the range for the possible MRs at each extreme efficient point by using the derivatives "to the right" and "to the left" of the DEA frontier at these points. To be specific, they compute these derivatives with the minimum and maximum multiplier ratios. Based on this approach, Prior and Surroca (2005) determine the MR as the minimum ratio for the corresponding weights in order to avoid infinite values. Since unbounded solutions can nevertheless exist in certain cases, they introduce a new set of constraints into the model of Rosen et al. (1998) in order to obtain the minimum ratios that ensure bounded weights. The MRs obtained by applying this method are used by these authors to determine strategic groups in an application to the Spanish banking industry.

4.6 Looking for Non-zero Weights

Zero weights are a source of trouble when estimating marginal rates. Zero weights also mean that some of the inputs and/or outputs initially considered for the analysis are eventually ignored. And zero weights, by duality, imply non-zero slacks in the primal envelopment formulation of the DEA model, which means that the unit under assessment is evaluated with reference to a point that is not on the Pareto-efficient frontier.

We deal with this issue in this section and review the literature on the multiplier bound approaches that have been proposed with the purpose of avoiding either zero weights or non-zero slacks. We also include here those that deal with the assessment of the DMUs that are on the weak efficient frontier or are projected onto it, i.e. those in $F \cup NF$ (in Bessent et al. 1988, these DMUs are referred to as "not naturally enveloped inefficient units"), since all these approaches eventually pursue the same objective.

First of all, it is to be noted that many of the approaches we have discussed in the previous sections may help mitigate the problem of zero weights and they can even lead to non-zero weights insofar as they involve reducing the weight flexibility.

Imposing a lower bound on the multipliers in the dual formulation of the DEA model is an easy way to avoid zero weights. In fact, this is the objective of the well-known non-Archimedean ε in the multiplier DEA models, but these models do not produce efficiency scores that can be readily used. To avoid this, some approaches replacing ε with a specific lower bound have been proposed. In the simplest case, a specific lower bound can be set, which ensures that no variable is eventually ignored in the assessments. A problem with this type of constraints is that the results provided depend on the specific value that is set for the bounds. For instance, in Sarrico et al. (1997), the constraints $u_r y_{r0} \geq 0.05$ are used for that purpose. Chang and Guh (1991) restrict the multipliers by means of a lower bound that is determined as the smallest non-zero value of the multipliers for all the variables of the unbounded DEA model. The main difficulty with this approach is that the resulting model may become infeasible [Chen et al. (2003) show how to modify this bound to avoid the infeasibility problems], together with the fact that the bounds are obtained following a procedure that does not consider the possible existence of alternate optima for the weights. Thus, different optimal solutions may lead to different bounds and these may lead to different efficiency scores for the units under assessment. Chen et al. (2003) develop an alternative multiplier bound approach in which the lower bounds are determined by SCSC solution pairs for extreme efficient DMUs. To be specific, they propose to select for each DMU_q in E a SCSC solution $\left(v_i^q, u_r^q\right)$ by using (4.7) and then assess the DMUs in $F \cup NF$ with (4.1) after incorporating the following constraints: $u_r \geq u_r^*,\ r \in R$ and $v_i \geq v_i^*,\ i \in I$, where $v_i^* = \min_{q \in E} \left\{v_i^q\right\}$ and $u_r^* = \min_{q \in E} \left\{u_r^q\right\}$, the sets R and I being $R = \{r / y_{rj}$ has non-zero slack value, $\text{DMU}_j \in F \cup NF\}$ and $I = \{i / x_{ij}$ has non-zero slack value, $\text{DMU}_j \in F \cup NF\}$. As acknowledged by the authors, the results obtained may obviously vary depending on the SCSC solution that is chosen. We also include here the approach in Dyson and Thanassoulis (1988) which, in the case of having a single input (or a single output), provides lower bounds for the output (input) weights by imposing the condition that it cannot be used less than some percentage of the average input level the DMU being assessed uses per unit of output, which is estimated by means of a regression analysis.

The two-step procedure in Ramón et al. (2010a) to assess relative efficiency without slacks also introduces some weight restrictions in the DEA model in order to avoid projections onto the weak efficient frontier. The specification of these constraints is based on a criterion that tries to avoid the extreme dissimilarity between the weights that is often found in practise. In the first step, for each DMU_{j_0} in E, the variable φ_{j_0} is maximized subject to the conditions $z_I \leq v_i \leq h_I, \quad i = 1, \ldots, m, \quad z_O \leq u_r \leq h_O, \quad r = 1, \ldots, s, \quad z_I/h_I \geq \varphi_{j_0}$ and $z_O/h_O \geq \varphi_{j_0}$, where v_i and u_r are the optimal solutions of the dual multiplier formulation of the CCR model for DMU_{j_0}, and z_I, h_I, z_O and h_O are non-negative variables. This problem looks for the least dissimilar weights that allow DMU_{j_0} to be rated as efficient. Then, this information provided by the extreme efficient units is summarized by means of the following scalar $\varphi^* = \min_{j \in E} \varphi_j^*$, which is used in the second step as the bound that specifies allowable differences in input multipliers and in output multipliers in the efficiency assessments. To be specific, in this second step, the efficiency of a given DMU_0 in $F \cup NF$ is assessed with the model resulting from adding to the dual multiplier formulation the following constraints: $z_I \leq v_i \leq h_I$, $i = 1, \ldots, m$, $z_O \leq u_r \leq h_O$, $r = 1, \ldots, s$, $z_I/h_I \geq \varphi^*$ and $z_O/h_O \geq \varphi^*$. As a consequence, with this model, we force the DMU_0 under evaluation to be assessed with reference to a set of weights that cannot be more dissimilar than those of the DMU in E that needs to unbalance its weights more (as measured by φ^*) in order to be rated as efficient. This procedure yields non-zero weights while at the same time it tries to avoid large differences in the multipliers as much as possible.

The existing work on facet models, which relates to the extension of the facets of the frontier, is in particular intended to address the problems with zero weights, and consequently with non-zero slacks. The basic idea is to project the not naturally enveloped inefficient units (those in $F \cup NF$) onto the extension of a given Pareto-efficient facet of the frontier. The existing approaches differ in the selection of the facet to be extended. We briefly describe here the approaches in Olesen and Petersen (1996) and in Portela and Thanassoulis (2006).

Olesen and Petersen (1996) show that avoiding non-zero weights does not guarantee well-defined rates of substitution along the estimated efficient frontier because this requires the existence of FDEFs. Note that if a given facet of the efficient frontier is not of full dimension, there will be infinitely many supporting hyperplanes containing it, each one associated with a different set of weights that will give rise to different estimations of the marginal rates of substitution. Their method seeks efficiency evaluations regarding FDEFs in order to guarantee well-defined rates of substitution. If FDEFs exist (which rarely happens in practise), then these authors define the following two technologies based on these facets: the extended facet PPS T_{EXFA}, as the intersection of half-spaces defined by FDEF-generating supporting hyperplanes and the non-negative orthant, and the FDEF PPS T_{FDEF}, as the union of the PPSs spanned by FDEFs, which satisfy

$T_{\text{FDEF}} \subseteq T_{\text{DEA}} \subseteq T_{\text{EXFA}}$(note that in contrast to T_{EXFA}, T_{FDEF} does not result from extending facets of T_{DEA} and may be non-convex). In these two new technologies, which are determined solely by the data, substitutional rates along the efficient frontier are well-defined. And thus, the efficiency scores computed relative to them provide a lower and an upper bound on the efficiency rating of the DMU under evaluation, respectively, the CCR score lying in between them. Obviously, non-zero weights can be guaranteed since only FDEFs are used.

Portela and Thanassoulis (2006) propose a procedure aimed at avoiding zero weights which also uses FDEFs. It is based on a two-step procedure that also guarantees projections onto facets where marginal rates of substitution are well-defined. In the first step, all the FDEFs are identified as well as their corresponding unique DEA optimal weights v_i^* and u_r^*. In the second step, the efficiency of DMU_0 is assessed with the model resulting of adding to the dual multiplier formulation the following AR constraints:

$$\min \frac{v_i^*}{v_1^*} \leq \frac{v_i}{v_1} \leq \max \frac{v_i^*}{v_1^*}, \quad i = 2, \ldots, m, \; \min \frac{u_r^*}{v_1^*} \leq \frac{u_r}{u_1} \leq \max \frac{u_r^*}{v_1^*}, \quad r = 2, \ldots, s, \tag{4.8}$$

where the weight v_1 is taken as the numeraire and the min and max are computed through the different FDEFs. It is to be noted that, in contrast to the method by Olesen and Petersen, this procedure allows to using non-observed facets, provided that these guarantee marginal rates that are no higher than the maximum observed neither lower than the minimum observed (an extension of this approach that only uses observed facets or their extensions is also developed).

For other facet model approaches, see Green et al (1996b), the "constrained facet analysis" (CFA) in Bessent et al. (1988) and the "controlled envelopment analysis" (CEA) in Lang et al. (1995).

Finally, it should be noted that the problem with the non-zero slacks has also been approached by dealing with the primal envelopment formulation of the DEA models. This is the case of the so-called "generalized efficiency measures" (GEMs) (see, e.g. Cooper et al. 1999; Pastor et al. 1999; and Tone 2001), which are efficiency measures especially designed to account for both radial and non-radial inefficiencies, and so, avoid the problems with the slacks.

4.7 Avoiding Large Differences in the Values of Multipliers

It has been claimed in the literature that there is a need to exercise some control over the variation in factor weights resulting from DEA flexibility. We can distinguish between two kinds of flexibilities in the DEA weights: On the one hand, large differences in the weights from variable to variable mean that in the assessment of the unit being evaluated the factors considered play very

different roles, which may be a concern. In the extreme case, when zero weights exist, the corresponding variables are ignored in the analysis. On the other hand, in some cases (and for certain purposes), it may be considered unacceptable that the same factor is accorded widely different weights when assessing different units (see Roll and Golany 1993; Pedraja-Chaparro et al. 1997; and Thanassoulis et al. 2004 for discussions).

Cook and Seiford (2008) state that "the AR concept was developed to prohibit large differences in the values of multipliers." Indeed, the approaches to the problems discussed in the first part of the chapter, which use not only AR constraints but also other types of weight restrictions, may help reduce differences in the weights attached both to the inputs and to the outputs since they restrict the weight flexibility.

Among the approaches aimed at avoiding differences in the weights within the optimal solutions of a given DMU (by only using the information in data), the procedure in Ramón et al. (2010a), to assess relative efficiency without slacks described in the previous section can also be seen as reducing weight dispersion since the value φ^* provides a limit for the allowable differences in the input weights and in the output weights so that if less dispersion were imposed then at least one efficient unit would become inefficient. The value φ^* is also used as a limit for the differences in weights in the assessment of the inefficient units, which means that these are assessed with reference to a set of weights that cannot be more dissimilar than those of the DMU in E that needs to unbalance more its weights in order to be rated as efficient.

Bal et al. (2008) propose to incorporate the minimization of the coefficient of variation (CV) for input–output weights into the CCR model in order to have more homogeneous weights as well as to improve discrimination. The model proposed, called CVDEA, results from replacing the objective in the dual multiplier formulation of the CCR model with the following non-linear one

$$\sum_{r=1}^{s} u_r y_{r0} - \frac{\sqrt{\sum_{r=1}^{s} (u_r - \bar{u})^2/(s-1)}}{\bar{u}} - \frac{\sqrt{\sum_{i=1}^{m} (v_i - \bar{v})^2/(m-1)}}{\bar{v}},$$

where $\bar{v}$ and $\bar{u}$ represent the arithmetic means of the input and output weights, respectively. Wang and Luo (2010) point out difficulties with the use of the CV in this model as this would be aggregating and averaging weights with different dimensions and units, which is meaningless.

The literature has also dealt with the differences in the DEA weights that are accorded to a given variable in the efficiency evaluation across the different DMUs. DEA provides unit-specific weights as opposed to the common set of weights (CSWs) traditionally used in all engineering, and most economic, efficiency analyses. This variation in weights may be partially justified by the different circumstances under which the DMUs operate, and which are not captured by the chosen set of inputs and outputs factors, whereas CSW is to be used when there is no need (nor wish) to allow for additional, individual

circumstances (see Roll et al. 1991 for discussions). The use of unit-specific weights is intended to show the unit under assessment in its best possible light; this, in particular, often produces a large number of efficient DMUs, which is avoided with a CSW. In fact, there are some DEA-like approaches aimed at providing a CSW which, in most cases, are based on the idea of minimizing the differences between the DEA efficiency scores and those obtained with the CSW. Note that the DEA efficiency scores are always greater than or equal to those obtained relative to a CSW. Roll et al. (1991) suggest that a possible meaning of the efficiencies computed with a CSW in the context of DEA is that such values represent the part of a DMU's performance which can be explained when assuming uniformity of circumstances. Let θ_j^* be the DEA efficiency score of a given DMU_j and $\theta_j(u,v) = \sum_{r=1}^{s} u_r y_{rj} / \sum_{i=1}^{m} v_i x_{ij}$ that obtained with the vectors of non-negative weights $v = (v_1,\ldots v_m)^{\mathrm{T}}$ and $u = (u_1, \ldots, u_s)^{\mathrm{T}}$. Then, $\theta_j^* \geq \theta_j(u,v), \quad \forall u, v$. The idea of the following approaches is to determine a couple of weight vectors u and v that minimize the deviations between the θ_j^*'s and the $\theta_j(u,v)$'s. Kao and Hung (2005) derive a family of CSWs by minimizing the generalized family of distance measures $D_p(u,v) = \left[\sum_{j=1}^{n} (\theta_j^* - \theta_j(u,v))^p\right]^{1/p}$, $p \geq 1$. This approach results from applying the compromise solution approach introduced by Yu (1973), the standard DEA efficiency scores θ_j^* regarding the ideal solution for the DMUs to achieve. Cook and Zhu (2007) also deal with these distances but relax the objective to groups of DMUs which operate in similar circumstances. See also Despotis (2002) and Liu and Peng (2008), which are described with more details in the next section, who also follow this idea in the specification of CSWs.

Other type of CSW approaches can be found in Roll and Golany (1993). There, the authors propose three models that are aimed at finding a CSW by (1) looking for central values of the range for each variable, (2) maximizing the average efficiency of all DMUs and (3) maximizing the number of efficient DMUs.

We note again that there are other approaches to finding a CSW that are not based on the use of DEA techniques. See Sinuany-Stern et al. (1994), Friedman and Sinuany-Stern (1997) and Sinuany-Stern and Friedman (1998) who use multivariate techniques.

Finally, it should be mentioned that there are also some intermediate approaches between DEA and CSW, which still permit weights to vary across units, but try to avoid large differences in the multipliers. In the context of cross-efficiency evaluation, Ramón et al. (2011) have proposed a peer-evaluation of units based on the profiles of weights that are more similar among themselves of the efficient DMUs. Other approaches allow the weights to vary but only within prescribed bounds. Roll et al. (1991) propose different techniques to setting weights bounds which, in all cases, rely on initial runs of unbounded models. For example, once the weight matrix is compiled, they propose to set bounds for each absolute weight by (1) either eliminating outliers or specifying that a certain percentage of weights falls within them and (2) specifying an acceptable ratio of

variation for each weight across units. They also propose as an alternative to start with some known set of common weights and specify allowable limits of variation around them. In the efficiency analysis of highway maintenance patrols in Cook et al. (1994), the different factors are accorded a level regarding their range (down or up) and a level for their acceptable flexibility (narrow, tight or wide) in order to make a choice from the range of weights obtained in the unbounded runs. To do it, considerations regarding the relative importance of the variables and the reliability of data were taken into account. Note that these approaches ignore the existence of alternate optima, which might be suggesting to a priori make a choice of weights among alternate optima according to some suitable criterion.

4.8 Improving Discrimination and Ranking Units

Poor discrimination is often found in the assessment of performance with DEA models, which means that many of the DMUs are classified as efficient or are rated near the maximum efficiency score. This can be a result of uniformity of performance between units but in practise it is mostly a consequence of having a small number of DMUs as compared to that of the inputs and outputs, although there may be other reasons such as the existence of subsets of units with an unusual mix of inputs and outputs or operating at a different scale size to the rest of units (this latter in case of using variable returns to scale (VRS)).

To improve discrimination, we can proceed in different manners: for example, we can increase the number of DMUs or reduce the number of inputs and/or outputs by eliminating or aggregating some of them. Nevertheless, sometimes the number of DMUs cannot be increased and/or it is wanted to maintain the specific variable selection. This is why different procedures have been developed with the aim of improving discrimination. Incorporating weight restrictions reduces the flexibility in weights and this generally improves discrimination as a side effect. In this sense, many of the approaches mentioned in this chapter can help improve discrimination. In this section, we only deal with approaches that are specifically developed to address this issue and that do not require prior information. See also the reviews on this problem in Adler et al. (2002), Angulo-Meza and Estellita Lins (2002) or Podinovski and Thanassoulis (2007).

Reducing the dispersion in weights as a way to restrict their flexibility is the idea behind some of the approaches developed to improve discrimination. In fact, the CVDEA model described in the previous section was developed with the aim of having available more reasonable/homogeneous input and output weights as well as improving discrimination. With these two purposes, Li and Reeves (1999) develop what they call multiple criteria data envelopment analysis (MCDEA) in the framework of multiple objective linear programming (MOLP). Given that maximizing $\sum_{r=1}^{s} u_r y_0$ in the CCR model produce insufficient discrimination, they introduce two other objective functions in the classical

Table 4.4 Efficiency scores for different non-dominated solutions of model (4.9)

	Efficiency scores					
DMU	Classical (C1)	C2	C3	C1 + C2 + C3	C1 + C2	C1 + 5 × C2
1	1	0.742	1	1	1	0.961
2	1	0.984	0.956	0.956	0.993	0.993
3	1	0.796	0.765	0.765	0.796	0.796
4	0.820	0.674	0.577	0.577	0.756	0.702
5	1	0.860	1	1	1	1
6	1	1	1	1	1	1
7	1	0.864	1	1	1	0.864

DEA model. Defining the deviation variables $d_j = \sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} u_r y_{rj}$ they solve the following model:

$$
\begin{aligned}
&\text{Min} \quad d_0 \quad (\text{or Max } \sum_{r=1}^{s} u_r y_0),\\
&\text{Min} \quad M,\\
&\text{Min} \quad \sum_{j=1}^{n} d_j,\\
&\text{s.t. :} \quad \sum_{i=1}^{m} v_i x_{i0} = 1,\\
&\quad -\sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} u_r y_{rj} + d_j = 0, \qquad \forall j,\\
&\quad M - d_j \geq 0, \qquad \forall j,\\
&\quad v_i,\ u_r,\ d_j \geq 0, \qquad \forall i, r, j.
\end{aligned}
\tag{4.9}
$$

The variable M in the second objective represents the maximum quantity among all deviation variables d_j while the third objective function is the deviation sum. The feasible region for the weights in this MOLP model is the same as that in the CCR model but, in contrast to the optimal solutions provided by the CCR model, (4.9) provides non-dominated solutions. To solve (4.9), we need to select the specific method to be used in order to provide a non-dominated solution. Once this is obtained, the efficiency score for the DMU_0 is computed as $1 - d_0$. (Note that values of d_0 can vary for the different non-dominated solutions). Obviously, if we attach some importance to the second and third objectives, we will be preventing DMU_0 from only minimizing d_0, which improves discrimination.

To illustrate the different results that can be obtained by using model (4.9), Table 4.4 shows the efficiency scores for the data in Table 4.1 corresponding to individually minimizing each of the three objectives (under C1, C2 and C3, respectively) as well as others such as the sum of the three criteria (C1 + C2 + C3), the sum of the two first criteria (C1 + C2) and the sum of first criterion plus five times

the second one (C1 + 5 × C2). We can see in this example that minimum criterion (C3), C1 + C2 + C3 and C1 + C2 reduce the number of efficient DMUs from six to four, C1 + 5 × C2 yields only two efficient units while minimax criterion (C2) is the most restrictive one leading to a unique efficient DMU.

The results reported in Li and Reeves (1999) show that this approach also yields more homogeneous weights.

Because of the complexity of multiple objective problems, Bal et al. (2010) propose to solve the above MCDEA model by using weighted goal programming. This allows them to convert (4.9) into a single objective problem called GPDEA which minimizes the sum of the unwanted deviations.

Other approaches use the idea of CSW, which obviously restricts the flexibility in weights and so it may help improve discrimination. The approach in Despotis (2002) follows the basic idea mentioned in the previous section of providing CSWs by minimizing the deviations between the DEA efficiency scores and those obtained with the common weights. To be specific, this author minimizes a convex combination of these deviations measured in terms of both D_1 and D_∞ distances, $t(1/n)D_1(u,v) + (1-t)D_\infty(u,v)$, where t is a parameter in $[0,\ 1]$ whose specification leads to a different CSW. Then, Despotis defines the globally efficient units as those that maintain their 100% efficiency score in the light of at least one CSW. This author states that considering only the globally efficient units drastically reduces the set of DEA efficient DMUs and the discriminating power of DEA is thus improved. He also shows how we can rank the DEA efficient units by using both the number of times that they maintain their efficiency scores under global assessments and their average global efficiency score.

Liu and Peng (2008) propose a method to determine a CSW for the efficiency evaluation of efficient units which is used to rank them. The CSW is obtained by minimizing the following sum of deviations $\sum_{j\in E}(\Delta_j^I + \Delta_j^O)$, where Δ_j^I and Δ_j^O are such that $\left(\sum_{r=1}^{s} u_r y_{rj} + \Delta_j^O\right) / \left(\sum_{i=1}^{m} v_i x_{ij} - \Delta_j^I\right) = 1$ and E represents the total set of efficient DMUs, either extreme or not. It is shown that this procedure is equivalent to maximizing $\sum_{r=1}^{s} u_r \sum_{j\in E} y_{rj} - \sum_{i=1}^{m} v_i \sum_{j\in E} x_{ij}$, subject to the usual constraints in the dual formulation of the CCR model but in this case only regarding the efficient units. This can be seen as a way of finding the weight vector that maximizes the efficiency of the aggregated DMU resulting from summing the value of the inputs and the outputs of the efficient units. The DEA efficient units are then ranked according to their absolute efficiency score calculated with the obtained CSW. An additional ranking rule is provided in order to further derive an ordering of the units rated 1 by this score. In fact, these two latter approaches develop specific procedures to rank the efficient DMUs.

Improving discrimination is intimately related to ranking units. We therefore now focus on the cross-efficiency evaluations as an approach to ranking the whole set of DMUs which is based essentially on the choice of weights that the DMUs make.

DEA models give weights that are unit-specific and, therefore, provide a self-evaluation of the DMUs which cannot be used to derive an ordering. The basic idea of cross-efficiency evaluation is to assess each unit with the weights of all the DMUs instead of with its own weights only. The cross-efficiency score

of a given unit is usually calculated as the average of its cross-efficiencies, which are obtained with the weight profiles provided by all the DMUs. Let $(v_1^{d*}, ..., v_m^{d*}, u_1^{d*}, ..., u_s^{d*})$ be an optimal solution of the multiplier formulation for DMU_d. Then, the cross-efficiency of a given DMU_j using the profile of weights provided by DMU_d is obtained as $E_{dj} = \sum_{r=1}^{s} u_r^{d*} y_{rj} \big/ \sum_{i=1}^{m} v_i^{d*} x_{ij}$, and the cross-efficiency score of DMU_j is defined as the average of these cross-efficiencies $\bar{E}_j = (1/n) \sum_{d=1}^{n} E_{dj}$, $j = 1, \ldots, n$. Thus, each unit is evaluated within the range of weights of all the DMUs. And, as a result, cross-efficiency scores provide a peer-evaluation instead of a self-evaluation and, consequently, can be used to rank the DMUs.

Cross-efficiency evaluations have been used in applications in different contexts: Sexton et al. (1986) to nursing homes, Oral et al. (1991) to R&D projects, Doyle and Green (1994b) to higher education, Green et al. (1996a) to preference voting, Chen (2002) to electricity distribution sector, Lu and Lo (2007) to economic environmental performance and Wu et al. (2009b) and Cooper et al. (2011) to sport, at the Summer Olympic and to ranking basketball players, respectively.

However, there are also some problems with cross-efficiency evaluations. Perhaps, the most important difficulty is the possible existence of alternate optima for the weights obtained when solving (4.2), which may lead to different cross-efficiency scores depending on the choice that is made. The use of alternative secondary goals to the choice of weights among the alternative optimal solutions has been suggested as a potential remedy to the possible influence of this difficulty which may reduce the usefulness of cross evaluations. The idea of most of the existing proposals along this line is to implicitly provide for each DMU_d a set of optimal weights obtained after imposing some condition on the resulting cross-efficiencies. That is the case of the well-known benevolent and aggressive formulations (Sexton et al. 1986; Doyle and Green 1994a). For instance, the benevolent formulation selects weights that maintain the self-efficiency score of the unit under assessment while enhancing the cross-efficiencies of the other DMUs as much as possible, whereas the aggressive formulation also maintains the self-efficiency score while diminishing the rest of cross-efficiencies. To be precise, in the case of the benevolent approach the profile of weights provided by DMU_d is the optimal solution of the following problem

$$
\begin{aligned}
\text{Max} \quad & \sum_{j=1}^{n} d_j, \\
\text{s.t.:} \quad & \sum_{i=1}^{m} v_i^d x_{id} = 1, \\
& \sum_{r=1}^{s} u_r^d y_{rd} = \theta_d^*, \\
& -\sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} u_r y_{rj} + d_j = 0, \quad j = 1, ..., n, \\
& v_i, u_r \geq 0, \qquad \forall i, r,
\end{aligned}
\tag{4.10}
$$

where θ_d^* is the DEA efficiency score of DMU_d (the aggressive formulation simply minimizes in (4.10) instead of maximizing). It should be noted that (4.10) does not actually maximize globally the resulting cross-efficiency scores, since this would lead to a non-linear model having in its objective a sum of ratios, which might be difficult to solve. Instead, the difference between the numerator and the denominator of these ratios is used as a surrogate (see Doyle and Green 1994a for alternative surrogates). See also Liang et al. (2008a) and Wang and Chin (2010b) for extensions of these models. We also include here the paper by Wu et al. (2009e) which, in the evaluation of a given unit, propose as a secondary goal to maximize the sum of the ranks of the remaining DMUs.

A different approach can be found in Wang and Chin (2010a) and in Ramón et al. (2010b), previously mentioned in this chapter, wherein the weights of each DMU are determined without considering their impact on the other DMUs. In the former paper, the choice of weights that each DMU makes seeks to maximize the relative contribution of its outputs, and in the latter the idea is to reduce differences in the weights that each DMU attaches both to the different inputs and to the different outputs. See also the related papers including game approaches: Wu et al. (2009c) and Liang et al. (2008b), which provide two approaches from the perspective of the cooperative and non-cooperative games respectively (this latter is extended to the variable returns to scale (VRS) case in Wu et al. (2009a)) and Wu et al. (2009d) as a bargaining game.

It has also been claimed in the literature that cross-efficiency may be useful to avoid unrealistic weighting schemes without the need to elicit weight restrictions, which require external information on prices/costs or expert knowledge. In fact, Anderson et al. (2002) think that cross evaluation appears to eliminate unrealistic weighting schemes and state that "since the cross-efficiency value is a function of all the weighting schemes, it has been proposed that the unrealistic schemes are, in effect, cancelled out." Ramón et al. (2010b) agree with this idea but also believe that we may have more comprehensive cross-efficiency scores if we avoid unreasonable weights in the average of cross-efficiencies. As a general approach, they propose to prevent unrealistic weighting schemes instead of expecting that their effects are eliminated in the cross evaluation, since the amalgamation of weights cannot guarantee that the unreasonable weights are eliminated. In particular, these authors call attention to the selection of weights made by the inefficient DMUs in cross-efficiency evaluations, which is an aspect of the analysis that has been hardly discussed in the literature. In practise, the inefficient DMUs generally have a unique optimal solution for their weights in the dual model, and what is more important, this solution usually has many zeros. Due to this uniqueness in cross-efficiency evaluations based on DEA weights, these DMUs can only choose this optimal solution as the weight profiles that they provide for the calculation of the cross-efficiencies. In fact, that is the case of many of the existing approaches, which includes obviously both the benevolent and the aggressive formulations. Since we usually have a large number of inefficient units as compared to that of efficient DMUs, we can conclude that, in these situations, one cannot expect that the effects of these

unrealistic weighting schemes are cancelled out in the summary that the cross evaluation makes. This is illustrated in the ranking of basketball players in Cooper et al. (2011), where the undesirable effects of the DEA unrealistic weighting schemes of the ineffective players, which in particular had many zeros, are shown: The results provided by this cross-efficiency evaluation have led to conclusions that are clearly inconsistent with the basketball expert opinion.

A possible way of sorting these problems out is suggested in Ramón et al. (2010b), where the idea of the proposed approach is to make a choice of weights for the efficient units according to a suitable criterion and to somehow reassess the inefficient DMUs trying to avoid the original DEA weights. To be specific, to do it they use the multiplier bound approach in Ramón et al. (2010a) briefly described in Sect. 4.6, which in particular guarantees non-zero weights even for the inefficient DMUs. Obviously, this is only one of the possibilities along this line that can be proposed, since other criteria to the choice of weights can be considered. A different route to deal with the problems with the weights of the inefficient units is that followed in the "peer-restricted" cross-efficiency evaluation in Ramón et al. (2011), where only the profiles of weights of the DMUs that have non-zero weights are considered, so we make a peer-evaluation of the units without the weights of some inefficient DMUs. This can be seen as an intermediate approach in between DEA, which provides a self-evaluation of each unit, and the standard cross-efficiency evaluation, which assesses each unit with the profiles of weights of all the DMUs. Aside from avoiding zero weights, the choice of weights made in that approach also seeks to reduce as much as possible the differences between the weights profiles selected.

Finally, we again point out that there are other alternatives to ranking units in DEA that are not based on DEA weights such as the use of the super-efficiency score (Andersen and Petersen 1993). See again the review in Adler et al. (2002) for ranking units.

4.9 Conclusions

The basic DEA models have been extended and enhanced with the aim of dealing with the difficulties and weaknesses of the original methodology. Many of these problems are associated with the specification of weights and have been addressed with some approaches that are eventually based on different choices and uses of DEA weights.

The specification of weights with DEA models has different appealing features such as the total flexibility in the choice of weights, which avoids the need to a priori know the values of the weights in the indexes of efficiency, or the fact that DEA provides weights that are unit-specific, which allows the different units to exploit their strengths in the efficiency assessments. However, these good properties are also an important source of trouble. On the one hand, as a consequence of the total weight flexibility, the results provided by the DEA

models are often inconsistent with the accepted views, we frequently find poor discrimination in the assessments of the different DMUs, the variables involved are attached very different weights or, in the extreme case, many of them are assigned a zero weight, which means that the corresponding inputs and/or outputs are eventually ignored in the analysis. And on the other, the use of unit-specific weights, which provides a self-evaluation of the DMUs, causes that we cannot derive an ordering of the DMUs.

We have also pointed out in this chapter that, in practise, the efficient units that we find are generally extreme efficient DMUs that have alternate optima for their weights and this, for certain purposes, raises the need to make a choice among them, whereas the inefficient units, which are usually on the weak efficient frontier or are projected onto it, have a unique solution for their weights and this usually has many zeros, which is a concern. This has been hardly discussed in the literature and is also a source of trouble as has been pointed out, for example, in the context of cross-efficiency evaluations.

Incorporating value judgements into the analysis by using weight restrictions in the DEA models usually helps mitigate the effects of all these problems, but their use requires prior knowledge or expert opinion. Based on the information provided by the data, some extensions and enhancements of the original methodology dealing with weights have also been developed, but we believe that more can still be done in this area. Some of them deal with the technology, as is the case of the approaches that use FDEFs of the frontier and define a new PPS or provide some weight bounds. These are often of help with the zero weights, the alternate optima for the weights or the estimation of marginal rates. However, FDEFs are rarely found in practise. Making a specific choice of weights according to some suitable criteria in the context of the problem(s) to be addressed is an approach that has been used in some of the methods that we have reviewed in this chapter. Maximizing the relative importance of the variables involved, reducing the dispersion in weights or even modifying the criterion of efficiency are simply some of those that have been mentioned in the literature. The use of such type of criteria could be explored in further developments of the DEA methodology.

As also noted in Cooper et al. (1996), the case of increasing returns to scale can be clarified by using Banker's most productive scale size to write $(X_o\alpha, Y_o\beta)$. The case $1 < \beta/\alpha$ means that all outputs are increased by at least the factor β and returns to scale are increasing as long as this condition holds. The case $1 > \beta/\alpha$ has the opposite meaning – viz., no output is increasing at a rate that exceeds the rate at which all inputs are increased. Only for constant returns to scale do we have $1 = \beta/\alpha$, in which case all outputs and all inputs are required to be increasing (or decreasing) at the same rate so no mix change is involved for the inputs.

The results in this chapter (as in the literature to date) are restricted to this class of cases. This leaves unattended a wide class of cases. One example involves the case where management interest is centred on only subsets of the outputs andinputs. A direct way to deal with this situation is to partition the inputs and outputs of interest and designate the conditions to be considered by

$(X_o^I\alpha, X_o^N, Y_o^I\beta, Y_o^N)$ where I designates the inputs and outputs that are of interest to management and N designates those which are not of interest (for such scale returns studies). Proceeding as described in the present chapter and treating X_o^N and Y_o^N as "exogenously fixed," in the spirit of Banker and Morey (1986), would make it possible to determine the situation for returns to scale with respect to the thus designated subsets. Other cases involve treatments with unit costs and prices as in Färe (1994) and Sueyoshi (1999).

The developments covered in this chapter have been confined to technical aspects of production. Our discussions follow a long-standing tradition in economics which distinguishes scale from mix changes by not allowing the latter to vary when scale changes are being considered. This permits the latter (i.e. scale changes) to be represented by a single scalar – hence the name. However, this can be far from actual practise, where scale and mix are likely to be varied simultaneously when determining the size and scope of an operation. See the comments by a steel industry consultant that are quoted in Cooper et al. (2000: 130) on the need for reformulating this separation between mix and scale changes in order to achieve results that more closely conform to needs and opportunities for use in actual practise.

There are, of course, many other aspects to be considered in treating returns to scale besides those attended to in the present chapter. Management efforts to maximize profits, even under conditions of certainty, require simultaneous determination of scale, scope and mix magnitudes with prices and costs known, as well as the achievement of the technical efficiency which is always to be achieved with *any* set of positive prices and costs. The topics treated in this chapter do not deal with such price-cost information. Moreover, the focus is on ex post facto analysis of already effected decisions. This can have many uses, especially in the control aspects of management where evaluations of performance are required. Left unattended in this chapter, and in much of the DEA literature, is the ex ante (planning) problem of how to use this knowledge in order to determine how to blend scale and scope with mix and other efficiency considerations when effecting future-oriented decisions.

Acknowledgments We are very grateful to Ministerio de Ciencia e Innovación (MTM2009-10479), the Generalitat Valenciana (ACOMP/2011/115) and to the IC2 Institute of The University of Texas at Austin for its financial support.

References

Adler N, Friedman L, Sinuany-Stern Z. Review of ranking in the data envelopment analysis context. Eur J Oper Res. 2002;140:249–65.

Allen R, Athanassopoulos A, Dyson RG, Thanassoulis E. Weights restrictions and value judgements in Data Envelopment Analysis: evolution, development and future directions. Ann Oper Res. 1997;73:13–34.

Andersen P, Petersen NC. A procedure for ranking efficient units in data envelopment analysis. Manag Sci. 1993;39:1261–4.
Anderson TR, Hollingsworth K, Inman LB. The fixed weighting nature of a cross-evaluation model. J Product Anal. 2002;18(1):249–55.
Angulo-Meza L, Estellita Lins MP. Review of methods for increasing discrimination in Data Envelopment Analysis. Ann Oper Res. 2002;116:225–42.
Bal H, Örkcü HH, Çelebioglu S. A new method based on the dispersion of weights in data envelopment analysis. Comput Ind Eng. 2008;54(3):502–12.
Bal H, Örkcü HH, Çelebioglu S. Improving the discrimination power and weights dispersion in the data envelopment analysis. Comput Oper Res. 2010;37(1):99–107.
Banker RD, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manag Sci. 1984;30:1078–92.
Banker RD, Morey RC. Efficiency analysis for exogenously fixed. Oper Res. 1986;34(4):513–21.
Beasley JE. Comparing University departments. Omega. 1990;18(2):171–83.
Bessent A, Bessent W, Elam J, Clark T. Efficiency frontier determination by constrained facet analysis. Oper Res. 1988;36(5):785–96.
Brockett PL, Charnes A, Cooper WW, Huang ZM, Sun DB. Data transformation in DEA cone ratio envelopment approaches for monitoring bank performances. Eur J Oper Res. 1997;98:250–68.
Bulla S, Cooper WW, Wilson D, Park KS. Evaluating efficiencies of turbo fan jet engines: a data envelopment analysis approach. J Eng Des. 2000;8(1):53–74.
Camanho AS, Dyson RG. Cost efficiency measurement with price uncertainty: a DEA application to bank branch assessments. Eur J Oper Res. 2005;161(2):432–46.
Chang KP, Guh YY. Linear production functions and Data Envelopment Analysis. Eur J Oper Res. 1991;52:215–23.
Charnes A, Cooper WW. Programming with linear fractional functionals. Naval Res Logist Quart. 1962;9:181–6.
Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res. 1978;2(6):429–44.
Charnes A, Cooper WW, Thrall RM. A structure for classifying and characterizing efficiency and inefficiency in Data Envelopment Analysis. J Product Anal. 1991;2:197–237.
Charnes A, Cooper WW, Huang ZM, Sun DB. Polyhedral Cone-Ratio DEA models with an illustrative application to large commercial banks. J Econ. 1990;46:73–91.
Chen TY. An assessment of technical efficiency and cross-efficiency in Taiwan's electricity distribution sector. Eur J Oper Res. 2002;137(2):421–33.
Chen Y, Morita H, Zhu J. Multiplier bounds in DEA via strong complementary slackness condition solutions. Int J Prod Econ. 2003;86:11–9.
Cook WD, Seiford LM. Data Envelopment Analysis (DEA) – thirty years on. Eur J Oper Res. 2008;192(1):1–17.
Cook WD, Zhu J. Within-group common weights in DEA: An analysis of power plant efficiency. Eur J Oper Res. 2007;178(1):207–16.
Cook WD, Kazakov A, Roll Y. On the measurement and monitoring relative efficiency of highway maintenance patrols. In: Charnes A, Cooper WW, Lewin AY, Seiford LM, editors. Data Envelopment Analysis. Theory, Methodology and Applications. Boston: Kluwer Academic; 1994.
Cooper WW, Thompson RG, Thrall RM. Introduction: extensions and new developments in DEA. Ann Oper Res. 1996;66:3–45.
Cooper WW, Park KS, Pastor JT. RAM: a range adjusted measure of inefficiency for use with additive models, and relations to other models and measures in DEA. J Product Anal. 1999;11 (1):5–42.
Cooper WW, Seiford LM, Tone, K. Data Envelopment Analysis: Comprehensive Text with Models, Applications, References and DEA-Solver Software. Kluwer Academic Publishers, Boston; 2000.

Cooper WW, Ramón N, Ruiz JL, Sirvent I. Avoiding large differences in weights in cross-efficiency evaluations: application to the ranking of basketball players. Journal of Centrum Cathedra, 2011.
Cooper WW, Ruiz JL, Sirvent I. Choosing weights from alternative optimal solutions of dual multiplier models in DEA. Eur J Oper Res. 2007;180(1):443–58.
Cooper WW, Ruiz JL, Sirvent I. Selecting non-zero weights to evaluate effectiveness of basketball players with DEA. Eur J Oper Res. 2009;195(2):563–74.
Despotis DK. Improving the discriminating power of DEA: focus on globally efficient units. J Oper Res Soc. 2002;53:314–23.
Doyle JR, Green RH. Efficiency and cross-efficiency in DEA: derivations, meanings and uses. J Oper Res Soc. 1994a;45(5):567–78.
Doyle JR, Green RH. Self and peer appraisal in Higher Education. High Educ. 1994b;28:241–64.
Dyson RG, Thanassoulis E. Reducing weight flexibility in data envelopment analysis. J Oper Res Soc. 1988;39:563–76.
Färe R, Grosskopf S, Lovell CAK. Production Frontiers. Cambridge University Press, Cambridge; 1994.
Friedman L, Sinuany-Stern Z. Scaling units via the canonical correlation analysis in the DEA context. Eur J Oper Res. 1997;100(3):629–37.
Green RH, Doyle JR, Cook WD. Preference voting and project ranking using DEA and cross-evaluation. Eur J Oper Res. 1996a;90:461–72.
Green RH, Doyle JR, Cook WD. Efficiency bounds in Data Envelopment Analysis. Eur J Oper Res. 1996b;89:482–90.
Joro T, Viitala EJ. Weight-restricted DEA in action: from expert opinions to mathematical models. J Oper Res Soc. 2004;55(8):814–21.
Kao C, Hung HT. Data envelopment analysis with common weights: the compromise solution approach. J Oper Res Soc. 2005;56(10):1196–203.
Khalili M, Camanho AS, Portela MCAS, Alirezaee MR. The measurement of relative efficiency using data envelopment analysis with assurance regions that link inputs and outputs. Eur J Oper Res. 2010;98:269–89.
Kuosmanen T, Post T. Measuring economic efficiency with incomplete price information: with an application to European commercial banks. Eur J Oper Res. 2001;134:43–58.
Lang P, Yolalan OR, Kettani O. Controlled envelopment by face extension in DEA. J Oper Res Soc. 1995;46(4):473–91.
Lee H, Park Y, Choi H. Comparative evaluation of performance of national R&D programs with heterogeneous objectives: a DEA approach. Eur J Oper Res. 2009;196(3):847–55.
Li XB, Reeves GR. Multiple criteria approach to data envelopment analysis. Eur J Oper Res. 1999;115(3):507–17.
Liang L, Wu J, Cook WD, Zhu J. Alternative secondary goals in DEA cross-efficiency evaluation. Int J Prod Econ. 2008a;113:1025–30.
Liang L, Wu J, Cook WD, Zhu J. The DEA game cross-efficiency model and its Nash equilibrium. Oper Res. 2008b;56(5):1278–88.
Liu FHF, Peng HH. Ranking of units on the DEA frontier with common weights. Comput Oper Res. 2008;35(5):1624–37.
Lozano S, Villa G, Guerrero F, Cortés P. Measuring the performance of nations at the Summer Olympics using data envelopment analysis. J Oper Res Soc. 2002;53(5):501–11.
Lu WM, Lo SF. A closer look at the economic-environmental disparities for regional development in China. Eur J Oper Res. 2007;183(2):882–94.
Olesen O, Petersen NC. Indicators of ill-conditioned data sets and model misspecification in Data Envelopment Analysis: An extended facet approach. Manag Sci. 1996;42:205–19.
Olesen OB, Petersen NC. Probabilistic bounds on the virtual multipliers in data envelopment analysis: polyhedral cone constraints. J Product Anal. 1999;12(2):103–33.
Olesen OB, Petersen NC. The use of data envelopment analysis with probabilistic assurance regions for measuring hospital efficiency. J Product Anal. 2002;17(1/2):83–109.

Oral M, Kettani O, Lang P. A mythology for collective evaluation and selection of industrial R&D projects. Manag Sci. 1991;37(7):871–85.
Paradi JC, Schaffnit C. Commercial branch performance evaluation and results communication in a Canadian bank - A DEA application. Eur J Oper Res. 2004;156(3):719–35.
Pastor JT, Ruiz JL, Sirvent I. An enhanced DEA Russell graph efficiency measure. Eur J Oper Res. 1999;115(3):596–607.
Pedraja-Chaparro F, Salinas-Jimenez J, Smith P. On the role of weight restrictions in Data Envelopment Analysis. J Product Anal. 1997;8:215–30.
Podinovski VV. Production trade-offs and weight restrictions in data envelopment analysis. J Oper Res Soc. 2004;55(12):1311–22.
Podinovski VV, Thanassoulis E. Improving discrimination in data envelopment analysis: some practical suggestions. J Product Anal. 2007;28(1–2):117–26.
Portela MCAS, Thanassoulis E. Zero weights and non-zero slacks: different solutions to the same problem. Ann Oper Res. 2006;145:129–47.
Prior D, Surroca J. Strategic groups based on marginal rates: an application to the Spanish banking industry. Eur J Oper Res. 2005;170(1):293–314.
Ramón N, Ruiz JL, Sirvent I. A multiplier bound approach to assess relative efficiency in DEA without slacks. Eur J Oper Res. 2010a;203(1):261–9.
Ramón N, Ruiz JL, Sirvent I. On the choices of weights profiles in cross-efficiency evaluations, Working paper (CIO-2010-5), Centro de Investigación Operativa, Universidad Miguel Hernández. Eur J Oper Res. 2010b;207(3):1564–72.
Ramón N, Ruiz JL, Sirvent I. Using data envelopment analysis to assess effectiveness of the processes at the university with performance indicators of quality. Int J Oper Quant Manag. 2010c;16(1):87–103.
Ramón N, Ruiz JL, Sirvent I. Reducing differences between profiles of weights: A "peer-restricted" cross-efficiency evaluation. Omega. 2011;39(6):634–41.
Ray SC, Seiford LM, Zhu J. Market entity behavior of Chinese state-owned enterprises. Omega. 1998;26(2):263–78.
Roll Y, Golany B. Alternate methods of treating factor weights in DEA. Omega. 1993;21 (1):99–109.
Roll Y, Cook WD, Golany B. Controlling factor weights in data envelopment analysis. IEEE Trans. 1991;23(1):2–9.
Rosen D, Schaffnit C, Paradi JC. Marginal rates and two-dimensional level curves in DEA. J Product Anal. 1998;9(3):205–32.
Saaty TL. The analytic hierarchy process. New York: McGraw-Hill International Book Company; 1980.
Sarrico CS, Dyson RG. Using DEA for planning in UK universities – an institutional perspective. J Oper Res Soc. 2000;51(7):789–800.
Sarrico CS, Dyson RG. Restricting virtual weights in data envelopment analysis. Eur J Oper Res. 2004;159:17–34.
Sarrico CS, Hogan SM, Dyson RG, Athanassopoulos AD. Data envelopment analysis and university selection. J Oper Res Soc. 1997;48(12):1163–77.
Schaffnit C, Rosen D, Paradi JC. Best practice analysis of bank branches: an application of DEA in a large Canadian bank. Eur J Oper Res. 1997;98:269–89.
Sexton TR, Silkman RH, Hogan AJ. Data envelopment analysis: critique and extensions. In: Silkman RH, editor. Measuring efficiency: an assessment of data envelopment analysis. San Francisco: Jossey-Bass; 1986. p. 73–105.
Shang J, Sueyoshi T. A unified framework for the selection of a flexible manufacturing system. Eur J Oper Res. 1995;85(2):297–315.
Shimshak DG, Lenard ML, Klimberg RK. Incorporating quality into data envelopment analysis of nursing home performance: a case study. Omega. 2009;37(3):672–85.
Sinuany-Stern Z, Friedman L. DEA and the discriminant analysis of ratios for ranking units. Eur J Oper Res. 1998;111(3):470–8.

Sinuany-Stern Z, Mehrez A, Barboy A. Academic departments efficiency via DEA. Comput Oper Res. 1994;21(5):543–56.
Sueyoshi T. DEA duality on returns to scale (RTS) in production and cost analyses: an occurrence of multiple solutions and differences between production-based and cost-based RTS estimates. Manag Sci. 1999;45(11):1593–1608.
Takamura Y, Tone K. A comparative site evaluation study for relocating Japanese government agencies out of Tokyo. Socio-Econ Plann Sci. 2003;37(2):85–102.
Thanassoulis E. Introduction to the theory and application of data envelopment analysis: a foundation text with integrated software. Boston: Kluwer Academic; 2001.
Thanassoulis E, Portela MCS, Allen R. Incorporating value judgements in DEA. In: Cooper WW, Seiford LW, Zhu J, editors. Handbook on Data Envelopment Analysis. Boston: Kluwer Academic; 2004.
Thompson RG, Brinkmann EJ, Dharmapala PS, González-Lima MD, Thrall RM. DEA/AR profit ratios and sensitivity of 100 large U.S. banks. Eur J Oper Res. 1997;98:213–29.
Thompson RG, Dharmapala PS, Thrall RM. Linked-cone DEA profit ratios and technical efficiency with application to Illinois Coal mines. Int J Prod Econ. 1995;39:99–115.
Thompson RG, Dharmapala PS, Rothenberg LJ, Thrall RM. DEA/AR efficiency and profitability of 14 major oil companies in US exploration and production. Comput Oper Res. 1996;23 (4):357–73.
Thompson RG, Lee E, Thrall RM. DEA/AR-efficiency of U.S. independent oil/gas producers over time. Comput Oper Res. 1992;19(5):377–91.
Thompson RG, Singleton FD, Thrall RM, Smith BA. Comparative site evaluations for locating a high-energy physics lab in Texas. Interfaces. 1986;16:35–49.
Tone K. A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res. 2001;130(3):498–509.
Wang YM, Chin KS. A neutral DEA model for cross-efficiency evaluation and its extension. Exp Syst Appl. 2010a;37:3666–75.
Wang YM, Chin, KS. Some alternative models for DEA cross-efficiency evaluation. Int J Prod Econ. 2010b;128(1):332–8.
Wang YM, Luo Y. A note on a new method based on the dispersion of weights in data envelopment analysis. Comput Ind Eng. 2010;56(4):1703–7.
Wong Y-HB, Beasley JE. Restricting weight flexibility in DEA. J Oper Res Soc. 1990;41:829–35.
Wu J, Liang L, Chen Y. DEA game cross-efficiency approach to Olympic rankings. Omega. 2009a;37:909–18.
Wu J, Liang L, Yang F. Achievement and benchmarking of countries at the Summer Olympics using cross-efficiency evaluation method. Eur J Oper Res. 2009b;197:722–30.
Wu J, Liang L, Yang F. Determination of the weights for the ultimate cross-efficiency using Shapley value in cooperative game. Exp Syst Appl. 2009c;36:872–6.
Wu J, Liang L, Yang F, Yan H. Bargaining game model in the evaluation of decision making units. Exp Syst Appl. 2009d;36:4357–62.
Wu J, Liang L, Yang F. Determination of cross-efficiency under the principle of ranking priority in cross-evaluation. Exp Syst Appl. 2009e;36:4826–9.
Yu PL. A class of solutions for group decision problems. Manag Sci. 1973;19:936–46.

Chapter 5
Malmquist Productivity Indexes and DEA

Rolf Färe, Shawna Grosskopf, and Dimitris Margaritis

Abstract In this chapter, we provide an overview of our recent work on data envelopment analysis (DEA) and Malmquist productivity indexes. First, we review the construction of static and dynamic DEA technologies. Based on these technologies we show how DEA can be used to estimate the Malmquist productivity index introduced by Caves et al. (Econometrica 50(6):1393–14, 1982) in the static case as well as its extension into the dynamic case.

Keywords DEA • Malmquist productivity index

5.1 Introduction

The Malmquist productivity change index introduced by Caves et al. (1982), (CCD), has become an important part of the DEA toolbox. Although introduced by CCD as a theoretical index defined in terms of distance functions, these distance functions have turned out to be very useful empirical tools.

The original work on data envelopment analysis (DEA) by Charnes et al. (1978) essentially shows how data-based activity analysis models can be solved using linear programming techniques to assess productive performance. Solutions to these problems are distance functions or equivalently Farrell (1957) measures of technical efficiency. Färe et al. (1989, 1994) made use of the connection between Farrell (1957), Charnes et al. (1978), and Caves et al. (1982) and introduced the aforementioned DEA estimation method for the Malmquist productivity index.

In Malmquist (1953), the original paper to which CCD refer and after whichthey named their proposed productivity index – Malmquist defined a quantity index

D. Margaritis (✉)
Department of Accounting & Finance, University of Auckland
Business School, Auckland, New Zealand
e-mail: d.margaritis@auckland.ac.nz

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_5,

as ratios of distances/distance functions in which observations were evaluated relative to an indifference curve, since he was working with consumer-based indexes. CCD substituted the technology frontier for the indifference curve to define a productivity index, in the spirit of Malmquist's consumer quantity index. Specifically, in the CCD definition of the output-oriented Malmquist productivity index, they used the output isoquant as the reference to which observations under evaluation were projected using an output distance function. Similarly, they chose an input isoquant as reference for the input-based Malmquist productivity index. The correspondent distance functions take value of unity if and only if the data belong to the respective isoquants.

There are two major surveys[1] of the Malmquist productivity index: Tone (2004) is directed toward the Operations Research (OR) audience, while Färe et al. (2008) are focused more toward economists. Here we take a middle road, following Tone (2004) and focusing our exposition with the OR audience in mind, while keeping reference with economics. A survey of this area is a moving target – we will be including recent work on dynamic Malmquist productivity indexes, endogenous technical change, as well as extensions based on alternate distance function specifications such as directional distance functions.

We begin by defining some DEA technologies, both static and dynamic. Next we discuss various projections to the frontiers of the DEA technologies including the familiar radial projections used in Farrell and Malmquist, which have natural associated dual value concepts well known to economists which we outline. We also discuss some nonradial projections such as the directional distance function, the Russell measure, and slacks-based measures. After these preliminaries we review the static Malmquist productivity index, its decomposition, provide links to traditional Solow and Törnqvist indexes, and also define the Luenberger productivity indicator defined in terms of directional distance functions. We close with the dynamic Malmquist productivity index.

5.2 DEA Technologies

The DEA technologies employed in this chapter are created from observations of decision-making units (hereafter DMUs). We denote their inputs by x, $x \in \Re_+^N$ and their outputs by y, $y \in \Re_+^M$. We assume that there are $k = 1, \ldots, K$ DMUs. The convex cone formed by these column vectors is the technology T, i.e.,

$$z_1[-x_{11} \cdots - x_{1N}\ y_{11} \ldots y_{1M}]' + \cdots + z_k[-x_{K1} \cdots - x_{KN}\ y_{K1} \ldots y_{KM}]' \geqq [-x_1 \ldots - x_N\ y_1 \ldots y_M]'$$

[1] See also Färe et al. (1998).

or in terms of summations

$$T = \left\{ (x,y) : \sum_{k=1}^{K} z_k x_{kn} \leqq x_n, \quad n = 1,\ldots,N. \right.$$
$$\sum_{k=1}^{K} z_k y_{km} \geqq y_m, \quad m = 1,\ldots,M$$
$$\left. z_k \geqq 0, \quad k = 1,\ldots,K \right\}, \tag{5.1}$$

where $z_k, k = 1,\ldots,K$ are the intensity variables which serve to form the convex combinations of data to form the technology set. The N and M inequalities impose strong disposability of inputs and outputs, respectively.

Specifically, if $(x,y) \in T$ and $x' \geqq x$ then $(x',y) \in T$ for input disposability and if $y' \leqq y$ then $(x,y') \in T$ for output disposability.

The nonnegativity constraints on the intensity variables, z_k, $k = 1, \ldots, K$ construct the technology as a cone, i.e., if $(x,y) \in T$ and $\lambda > 0$ then $((\lambda x, \lambda y) \in T)$, i.e., it satisfies constant returns to scale. Convexity of T follows, since one may take

$$\sum_{k=1}^{K} z_k = 1, \quad z_k \geqq 0, \quad k = 1,\ldots,K. \tag{5.2}$$

The convexity constraint together with the disposability conditions impose variable returns to scale on T. If we substitute

$$\sum_{k=1}^{K} z_k \leqq 1 \tag{5.3}$$

for the convexity constraint, then the technology T satisfies nonincreasing returns to scale rather than variable returns.

More fundamentally, we would like to know if there are constraints on the data vectors themselves that are required to ensure solutions to our linear programming problems. We appeal to Kemeny et al. (1956) who established that the following constraints provide the necessary conditions for the technology to be closed:

$$\sum_{k=1}^{K} x_{kn} > 0, \quad n = 1,\ldots,N$$
$$\sum_{n=1}^{N} x_{kn} > 0, \quad k = 1,\ldots,K$$
$$\sum_{k=1}^{K} y_{kn} > 0, \quad m = 1,\ldots,M$$
$$\sum_{m=1}^{M} y_{kn} > 0, \quad k = 1,\ldots,K. \tag{5.4}$$

Note that this implies that the x and y vectors are *not* required to be strictly positive, but rather nonnegative. Each row and column must contain at least one positive value, but otherwise zero values are allowed. Given these constraints on the data, T is closed and the output sets

$$P(x) = \{y : (x, y) \in T\}, \quad x \in \Re_+^N \tag{5.5}$$

are bounded. Given that T is closed it follows that $P(x)$ is closed and so are the input sets

$$L(y) = \{x : (x, y) \in T\}, \quad y \in \Re_+^M. \tag{5.6}$$

In terms of the data, the output and input sets take the form

$$P(x) = \left\{ y : \sum_{k=1}^{K} z_k x_{kn} \leqq x_n, \quad n = 1, \ldots, N. \right.$$
$$\sum_{k=1}^{K} z_k y_{km} \geqq y_m, \quad m = 1, \ldots, M$$
$$\left. z_k \geqq 0, \quad k = 1, \ldots, K \right\} \tag{5.7}$$

and

$$L(y) = \left\{ x : \sum_{k=1}^{K} z_k x_{kn} \leqq x_n, \quad n = 1, \ldots, N. \right.$$
$$\sum_{k=1}^{K} z_k y_{km} \geqq y_m, \quad m = 1, \ldots, M$$
$$\left. z_k \geqq 0, \quad k = 1, \ldots, K \right\}, \tag{5.8}$$

respectively.

To introduce our dynamic DEA model we need to account for time and the dynamic interaction between periods. Thus, denote time by t and let the use of output y^t be allocated toward final output, ${}^f y^t$ and intermediate output ${}^i y^t$ so that

$$y^t = {}^f y^t + {}^i y^t. \tag{5.9}$$

Hence, the trade-off is between "consuming" today and "investing" for tomorrow, just like a Ramsey (1928) model.

The following Fig. 5.1 illustrates the basic dynamic model.[2] Assume there are three time periods $t - 1$, t, and $t + 1$ with three production technologies expressed by their output sets, as shown in Fig. 5.1.

[2] This figure is adapted from Färe and Grosskopf (1996).

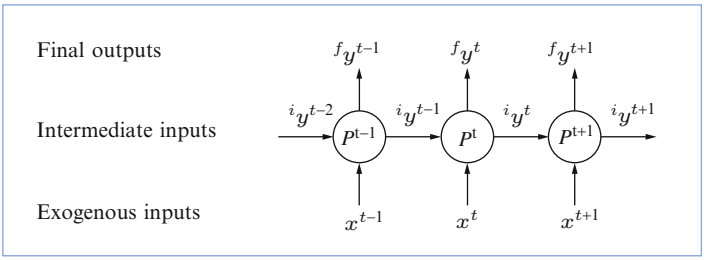


Fig. 5.1 The basic dynamic model

One DEA interpretation of the model may be expressed as[3]

$$
\begin{aligned}
P\left(x^{t-1},x^{t},x^{t+1},{}^{i}y^{t-2}\right) = \{&\left({}^{f}y^{t-1},{}^{f}y^{t},\left({}^{f}y^{t+1}+{}^{i}y^{t+1}\right)\right):\\
&{}^{f}y_{m}^{t-1}+{}^{i}y_{m}^{t-1}\leqq\sum_{k=1}^{K}z_{k}^{t-1}\left({}^{f}y_{km}^{t-1}+{}^{i}y_{km}^{t-1}\right),\quad m=1,\ldots,M\\
&\sum_{k=1}^{K}z_{k}^{t-1}x_{kn}^{t-1}\leqq x_{n}^{t-1},\quad n=1,\ldots,N\\
&\sum_{k=1}^{K}z_{k}^{t-1}\left({}^{i}y_{km}^{t-1}\right)\leqq{}^{i}y_{m}^{t-2},\quad m=1,\ldots,M\\
&z_{k}^{t-1}\geqq 0,\quad k=1,\ldots,K\\
&{}^{f}y_{m}^{t}+{}^{i}y_{m}^{t}\leqq\sum_{k=1}^{K}z_{k}^{t}\left({}^{f}y_{km}^{t}+{}^{i}y_{km}^{t}\right),\quad m=1,\ldots,M\\
&\sum_{k=1}^{K}z_{k}^{t}x_{kn}^{t}\leqq x_{n}^{t},\quad n=1,\ldots,N\\
&\sum_{k=1}^{K}z_{k}^{t}\left({}^{i}y_{km}^{t}\right)\leqq{}^{i}y_{m}^{t-1},\quad m=1,\ldots,M\\
&z_{k}^{t}\geqq 0,\quad \partial k=1,\ldots,K\\
&{}^{f}y_{m}^{t+1}+{}^{i}y_{m}^{t+1}\leqq\sum_{k=1}^{K}z_{k}^{t+1}\left({}^{f}y_{km}^{t+1}+{}^{i}y_{km}^{t+1}\right),\quad m=1,\ldots,M\\
&\sum_{k=1}^{K}z_{k}^{t+1}x_{kn}^{t+1}\leqq x_{n}^{t+1},\quad n=1,\ldots,N\\
&\sum_{k=1}^{K}z_{k}^{t+1}\left({}^{i}y_{km}^{t+1}\right)\leqq{}^{i}y_{m}^{t},\quad m=1,\ldots,M\\
&z_{k}^{t+1}\geqq 0,\quad k=1,\ldots,K\}.
\end{aligned}
\tag{5.10}
$$

[3] This model is also adapted from Färe and Grosskopf (1996). For an empirical application of this model, see Bogetoft et al. (2009).

The properties, like disposability and returns to scale, of this model follow from the static model. The computation of efficiency scores when applying the dynamic DEA model requires that initial conditions are specified, i.e., ${}^{i}y^{t-2}$, as well as the transversality conditions.

A second DEA interpretation is to let the investment enter into the column vectors and interpret investment as technological change.[4] The model takes the form

$$
\begin{aligned}
{}^{f}y_m^{t-1}+{}^{i}y_m^{t-1} &\leqq \sum_{k=1}^{K} z_k^{t-1}\left({}^{f}y_{km}^{t-1}+{}^{i}y_{km}^{t-1}+{}^{i}y_m^{t-2}\right), \quad m=1,\ldots,M\\
\sum_{k=1}^{K} z_k^{t-1}x_{kn}^{t-1} &\leqq x_n^{t-1}, \quad n=1,\ldots,N\\
z_k^{t-1} &\geqq 0, \quad k=1,\ldots,K\\
{}^{f}y_m^{t}+{}^{i}y_m^{t} &\leqq \sum_{k=1}^{K} z_k^{t}\left({}^{f}y_{km}^{t}+{}^{i}y_{km}^{t}+{}^{i}y_m^{t-1}\right), \quad m=1,\ldots,M\\
\sum_{k=1}^{K} z_k^{t}x_{kn}^{t} &\leqq x_n^{t}, \quad n=1,\ldots,N\\
z_k^{t} &\geqq 0, \quad k=1,\ldots,K\\
{}^{f}y_m^{t+1}+{}^{i}y_m^{t+1} &\leqq \sum_{k=1}^{K} z_k^{t+1}\left({}^{f}y_{km}^{t+1}+{}^{i}y_{km}^{t+1}+{}^{i}y_m^{t}\right), \quad m=1,\ldots,M\\
\sum_{k=1}^{K} z_k^{t+1}x_{kn}^{t+1} &\leqq x_n^{t+1}, \quad n=1,\ldots,N\\
\sum_{k=1}^{K} z_k^{t+1}\left({}^{i}y_{km}^{t+1}\right) &\leqq {}^{i}y_m^{t}, \quad m=1,\ldots,M\\
z_k^{t+1} &\geqq 0, \quad k=1,\ldots,K. \qquad (5.11)
\end{aligned}
$$

This is the model we will pursue further in this chapter. Note that if the investment variables ${}^{i}y_m^t=0$, then the dynamic model reduces to three static models, since we have that $y_{km}^{\tau}={}^{f}y_{km}^{\tau}+{}^{i}y_{km}^{\tau}$, $\tau=t-1,t,t+1$ which is denoted just by y_{km} above.

5.3 Projecting onto the Frontier

In building his quantity index, Malmquist radially projected the "observed" vector onto the frontier consisting of a fixed isoquant (indifference curve). Caves et al. (1982) adapted Malmquist's idea and projected radially onto the frontier

[4] This idea was developed by Färe et al. (2009).

consisting of the technology boundary. In the output direction they adapted the Shephard (1970) output distance function and in the input direction they applied the Shephard (1953) input distance function.

To estimate the output distance function relative to the static output set one can use linear programming (LP) to find the solution to the following problem for DMU k'

$$\left(D_o\left(x^{k'},y^{k'}\right)\right)^{-1}=\max\left\{\lambda:\sum_{k=1}^{K}z_kx_{kn}\leqq x_{k'n},\quad n=1,\ldots,N.\right.$$
$$\sum_{k=1}^{K}z_ky_{km}\geqq\lambda y_{k'm},\quad m=1,\ldots,M$$
$$\left.z_k\geqq 0,\quad k=1,\ldots,K\right\}. \tag{5.12}$$

The input-oriented model is estimated by

$$\left(D_i\left(y^{k'},x^{k'}\right)\right)^{-1}=\min\left\{\lambda:\sum_{k=1}^{K}z_kx_{kn}\leqq\lambda x_{k'n},\quad n=1,\ldots,N.\right.$$
$$\sum_{k=1}^{K}z_ky_{km}\geqq y_{k'm},\quad m=1,\ldots,M$$
$$\left.z_k\geqq 0,\quad k=1,\ldots,K\right\}. \tag{5.13}$$

Note that under constant returns to scale, i.e., when the intensity variables z_k, $k=1,\ldots,K$ are just restricted to be nonnegative, then

$$D_o\left(x^{k'},y^{k'}\right)=\frac{1}{D_i\left(y^{k'},x^{k'}\right)}, \tag{5.14}$$

i.e., they are reciprocal.

An important feature of the radial measures is that they have natural duals and hence are by nature suitable for Farrell (1957) type decompositions. To verify this let $p\in\Re_+^M$ denote output prices (which may vary across DMUs) and define the revenue function as

$$R\left(x^{k'},p\right)=\max\left\{py:\sum_{k=1}^{K}z_kx_{kn}\leqq x_{k'n},\quad n=1,\ldots,N.\right.$$
$$\sum_{k=1}^{K}z_ky_{km}\geqq y_m,\quad m=1,\ldots,M$$
$$\left.z_k\geqq 0,\quad k=1,\ldots,K\right\} \tag{5.15}$$

then

$$R\left(x^{k'},p\right)\geqq py \quad \text{for all} \quad y\in P\left(x^{k'}\right), \tag{5.16}$$

hence

$$R\left(x^{k'},p\right)\geqq \frac{py^{k'}}{D_o(x^{k'},y^{k'})} \tag{5.17}$$

and

$$\frac{R(x^{k'},p)}{py^{k'}}\geqq \frac{1}{D_o(x^{k'},y^{k'})} \tag{5.18}$$

where $R(x^{k'},p)/py^{k'}$ is revenue efficiency and $1/D_o\left(x^{k'},y^{k'}\right)$ is output efficiency.

Closing the inequality by multiplying with allocative efficiency AE_o yields the Farrell decomposition of revenue efficiency, namely

$$\frac{R(x^{k'},p)}{py^{k'}}=\frac{1}{D_o(x^{k'},y^{k'})}AE_o. \tag{5.19}$$

For the cost decomposition denote input prices by $w\in\Re^N_+$ (again, these may vary across DMUs) and define the cost function for DMU k' by

$$C\left(y^{k'},w\right)=\min\left\{wx:\sum_{k=1}^{K}z_kx_{kn}\leqq x_n,\quad n=1,\ldots,N.\right.$$
$$\sum_{k=1}^{K}z_ky_{km}\geqq y_{k'm},\quad m=1,\ldots,M$$
$$\left.z_k\geqq 0,\quad k=1,\ldots,K\right\} \tag{5.20}$$

then

$$C\left(y^{k'},w\right)\leqq wx \quad \text{for all} \quad x\in L\left(y^{k'}\right) \tag{5.21}$$

hence

$$\frac{C\left(y^{k'},w\right)}{wx^{k'}}\leqq \frac{1}{D_i\left(y^{k'},x^{k'}\right)} \tag{5.22}$$

and by multiplying with allocative inefficiency AE_i;

$$\frac{C\left(y^{k'},w\right)}{wx^{k'}}=\frac{1}{D_i\left(y^{k'},x^{k'}\right)}AE_i \tag{5.23}$$

the Farrell (1957) decomposition of cost efficiency into technical ($1/D_i$) and allocative (AE_i) efficiency is obtained.

Another function with a natural dual is the directional or shortage function. Let g_x and g_y be the directional input and output vectors[5] (the direction vector in which inputs are contracted and outputs expanded) then the directional distance function is defined in terms of the DEA model as

$$\vec{D}_t\left(x^{k'}, y^{k'}; g_x, g_y\right) = \max\left\{\beta : \sum_{k=1}^{K} z_k x_{kn} \leqq x_{k'n} - \beta g_{x_n}, \quad n = 1, \ldots, N. \right.$$
$$\sum_{k=1}^{K} z_k y_{km} \geqq y_{k'm} + \beta g_{y_m}, \quad m = 1, \ldots, M$$
$$\left. z_k \geqq 0, \quad k = 1, \ldots, K \right\}. \tag{5.24}$$

Note that if $g_x = 0$ and $g_y = y_{k'}$ then

$$\vec{D}_t\left(x^{k'}, y^{k'}; 0, y^{k'}\right) = \frac{1}{D_o(x^{k'}, y^{k'})} - 1 \tag{5.25}$$

and if $g_y = 0$and $g_x = x^{k'}$ then

$$\vec{D}_t\left(x^{k'}, y^{k'}; x^{k'}, 0\right) = 1 - \frac{1}{D_i(y^{k'}, x^{k'})}. \tag{5.26}$$

The dual to the directional distance function is the profit function. Let $p \in \Re_+^M$ be an output price vector and $w \in \Re_+^N$ be an input price vector. Since profit is zero under constant returns to scale, let us add the nonincreasing returns to scale constraint, $\sum_{k=1}^{K} z_k \leqq 1$, to the technology. Then the profit function is defined as

$$\pi(p, w) = \max\left\{py - wx : \sum_{K=1}^{K} z_k x_{kn} \leqq x_n, \quad n = 1, \ldots, N \right.$$
$$\sum_{K=1}^{K} z_k y_{km} \geqq y_m, \quad m = 1, \ldots, M$$
$$z_k \geqq 0, \quad k = 1, \ldots, K$$
$$\left. \sum_{k=1}^{K} z_k \leqq 1 \right\}. \tag{5.27}$$

[5] Note that each component is equipped with a unit of measurement.

This means that $\pi(p,w) \geqq py - wx$ for all (x, y) in the "nonincreasing" returns to scale technology above.

Thus, if the directional distance function is defined on the same technology we have

$$\pi(p,w) \geqq p\left(y + \vec{D}_T(\cdot) g_y\right) - w\left(x - \vec{D}_T(\cdot) g_x\right) \tag{5.28}$$

and

$$\frac{\pi(p,w) - (py - wx)}{pg_y + wg_x} \geqq \vec{D}_T(x, y; g_x, g_y). \tag{5.29}$$

By adding allocative efficiency $(\overrightarrow{AE})$ to this expression we obtain the decomposition

$$\frac{\pi(p,w) - (py - wx)}{pg_y + wg_x} = \vec{D}_T\left(x, y; g_x, g_y\right) + \overrightarrow{AE}. \tag{5.30}$$

This is the decomposition of the Nerlovian efficiency measure

$$\frac{\pi(p,w) - (py - wx)}{pg_y + wg_x} \tag{5.31}$$

into a technical component $\vec{D}_T(x, y; g_x, g_y)$ and an allocative component $\overrightarrow{AE}$.

There are other nonradial efficiency measures that project onto efficiency frontier. Starting with the Russell measure most of them project onto a subset of the frontier. In what follows we restrict our discussion to the study of input contractions.

The set we focus on is the subset of efficient input vectors. It is defined as

$$\text{Eff } L(y) = \{x \in L(y) : x' \leq x, x' \neq x, x' \notin L(y)\}. \tag{5.32}$$

In the activity or DEA model, this set is bounded for every $y \in \Re_+^M$ and since the input set $L(y)$ is closed, its closure must satisfy

$$\overline{\text{Eff } L(y)} \subseteq L(y), \tag{5.33}$$

and $\overline{\text{Eff } L(y)}$is compact.[6]

To find a model that unifies most input efficient measures that project onto the efficient subset, we generalize the directional distance function as follows. First, let $g_y = 0$ so that only input contraction is considered; second, introduce a vector of $\beta_i s$, namely, $\beta = (\beta_1, \ldots, \beta_N)$. Third, alter the objective function to read as

$$\beta_1 + \beta_2 + \cdots + \beta_N \tag{5.34}$$

[6] This allows us to cost minimize with nonnegative prices.

with $\beta_n \geq 0$.
Finally, assume that each $g_{x_n}, \quad n = 1, \ldots, N$ is strictly positive, then define

$$\mathrm{EM}_o = \max \sum_{n=1}^{N} \beta_n \tag{5.35}$$

$$\begin{aligned} \text{s.t.} \sum_{k=1}^{K} z_k x_{kn} &\leqq x_{k'n} - \beta_n g_{x_n}, \quad n = 1, \ldots, N \\ \sum_{k=1}^{K} z_k y_{km} &\leqq y_{k'm}, \quad m = 1, \ldots, M, \\ z_k &\geqq 0, \quad k = 1, \ldots, K. \end{aligned}$$

One may prove that[7]

$$\mathrm{EM}_o = 0 \Leftrightarrow x^{k'} \in \mathrm{Eff}\, L\left(y^{k'}\right). \tag{5.36}$$

Recall that each g_{x_n} has a unit of measurement so that β_n is independent of units of measurement, hence the βs can be added, preserving the independence condition.

In their slack-based measure of efficiency, Färe and Grosskopf (2010) took $g_{x_n} = 1$ for each $n = 1, \ldots, N$. By restricting to inputs, Tone (2001) modeled his efficiency measure through slacks. If we take

$$s_n = \beta_n g_{x_n} \quad \text{with} \quad g_{x_n} = x_{k'n}, \tag{5.37}$$

as the observed input, then a variation of Tone's model follows, namely,

$$\begin{aligned} &\max \sum_{n=1}^{N} \frac{s_n}{x_{k'n}} \\ &\text{s.t.} \sum_{k=1}^{K} z_k x_{kn} \leqq x_{k'n} - s_n, \quad n = 1, \ldots, N \\ &\sum_{k=1}^{K} z_k y_{km} \geqq y_{k'm}, \quad m = 1, \ldots, M \\ &z_k \geqq 0, \quad k = 1, \ldots, K. \end{aligned} \tag{5.38}$$

[7] The proof is similar to that of Färe and Lovell (1978) on the Russell measure.

Finally, if we take $g_{x_n} = x_{k'n}$, a variation of the Russell measure[8] is obtained as

$$\begin{aligned} &\min \sum_{n=1}^{N} \gamma_n \\ &\text{s.t.} \sum_{k=1}^{K} z_k x_{kn} \leqq x_{k'n}\gamma_n, \quad n = 1, \ldots, N \\ &\quad \sum_{k=1}^{K} z_k y_{km} \geqq y_{k'm}, \quad m = 1, \ldots, M \\ &\quad z_k \geqq 0, \quad k = 1, \ldots, K. \end{aligned} \tag{5.39}$$

where $\gamma_n = (1 - \beta_n)$.

Having said this, in what follows we restrict our discussion of the Malmquist productivity index to Shephard's output distance function. But we keep in mind that other projections onto the frontier may also be applied.

5.4 Productivity Indexes

Before introducing the Malmquist productivity index, let us begin with the simplest case: a world where one input is used to produce one output in two periods t and $t + 1$. In this case, the level of productivity is the ratio of output to input or average product

$$y^t/x^t \quad \text{and} \quad y^{t+1}/x^{t+1}. \tag{5.40}$$

The corresponding productivity change is the ratio of these average products

$$\frac{y^{t+1}/x^{t+1}}{y^t/x^t}. \tag{5.41}$$

Productivity has increased between t and $t + 1$ if this ratio is greater than unity.

The Malmquist productivity change index is formulated in terms of distance functions, which ultimately will allow us to include multiple inputs and outputs. To see this we can rewrite the relationship above as follows:

$$\frac{y^{t+1}/x^{t+1}}{y^t/x^t} = \frac{(y^{t+1}/x^{t+1})D_o(1,1)}{(y^t/x^t)D_o(1,1)} = \frac{D_o(x^{t+1}, y^{t+1})}{D_o(x^t, y^t)} \tag{5.42}$$

[8] See Färe and Lovell (1978).

given that the technology satisfies constant returns to scale since then

$$D_o(\lambda x, y) = \lambda D_o(x, y), \quad \lambda > 0, \tag{5.43}$$

and by definition we also have that

$$D_o(x, \lambda y) = \lambda D_o(x, y), \quad \lambda > 0. \tag{5.44}$$

Caves et al. (1982) defined two Malmquist (output-oriented) productivity change indexes, one for the reference technology P^t and another for P^{t+1}, or in terms of distance functions

$$M_o^t = D_o^t\left(x^{t+1}, y^{t+1}\right)/D_o^t(x^t, y^t) \tag{5.45}$$

and

$$M_o^{t+1} = D_o^{t+1}\left(x^{t+1}, y^{t+1}\right)/D_o^{t+1}(x^t, y^t). \tag{5.46}$$

One can show that if these productivity change indexes are equal, i.e.,[9]

$$D_o^t\left(x^{t+1}, y^{t+1}\right)/D_o^t(x^t, y^t) = D_o^{t+1}\left(x^{t+1}, y^{t+1}\right)/D_o^{t+1}(x^t, y^t) \tag{5.47}$$

then and only then is the technology Hicks output neutral. In terms of the output distance function, this is equivalent to

$$D_o^t(x, y) = A(t) D_o(x, y). \tag{5.48}$$

This is a rather restrictive assumption, which we wish to avoid. Instead, we take the geometric mean of the two indexes above, namely[10]

$$\begin{aligned} M_o\left(x^t, y^t, x^{t+1}, y^{t+1}\right) &= \left(M_o^t \cdot M_o^{t+1}\right)^{1/2} \\ &= \left(\frac{D_o^t(x^{t+1}, y^{t+1})}{D_o^t(x^t, y^t)} \frac{D_o^{t+1}(x^{t+1}, y^{t+1})}{D_o^{t+1}(x^t, y^t)}\right)^{1/2}. \end{aligned} \tag{5.49}$$

The index is illustrated in Fig. 5.2.[11] The figure illustrates the simple case of one input producing one output, and two technologies T^t and T^{t+1}, one for each period. In the DEA world, these technologies are based on observed data.

[9] See Färe et al. (1998).

[10] See Färe and Grosskopf (1992).

[11] This figure is adapted from Färe and Grosskopf (1996).

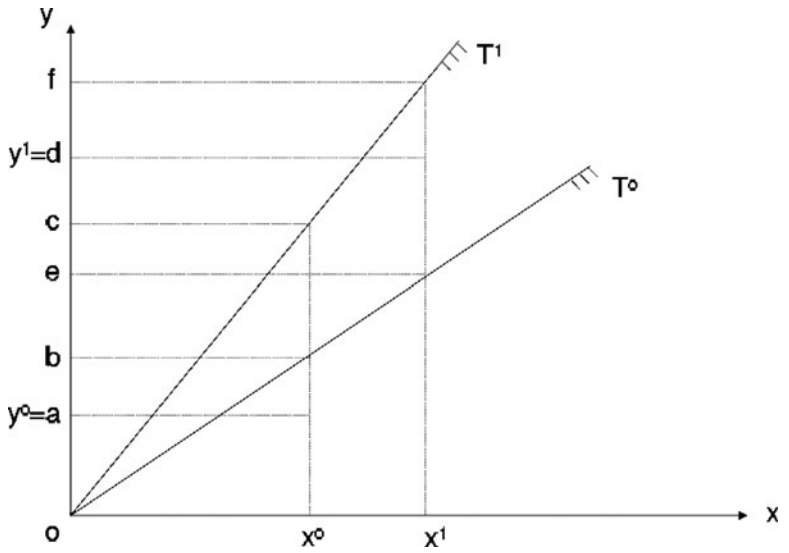


Fig. 5.2 Malmquist productivity change

In the figure the two observations (x^t, y^t) and (x^{t+1}, y^{t+1}) belong to the corresponding technologies, i.e., T^t and T^{t+1}. The t period observation is also feasible at $t + 1$, but (x^{t+1}, y^{t+1}) does not belong to T^t. In terms of distances along the output axis, the index equals

$$M_o\left(x^t, y^t, x^{t+1}, y^{t+1}\right) = \left(\frac{0c/0a}{0f/0b}\,\frac{0c/0d}{0f/0e}\right)^{1/2}. \tag{5.50}$$

Note that this may also be written as

$$M_o\left(x^t, y^t, x^{t+1}, y^{t+1}\right) = \left(\frac{0c}{0a}\,\frac{0e}{0f}\right)\left(\frac{0a}{0d}\,\frac{0b}{0e}\right)^{1/2}, \tag{5.51}$$

where the first expression in parentheses captures the change in technical efficiency between period t and $t + 1$ and the second measures the shift in the frontier of technology. In general, the Malmquist productivity change index decomposes multiplicatively into an efficiency change component (EFFCH) and a technical change component (TECH)[12]:

$$\text{EFFCH} = D_o^{t+1}\left(x^{t+1}, y^{t+1}\right)/D_o^t(x^t, y^t) \tag{5.52}$$

[12] We discuss alternative decompositions and interpretations presently.

and

$$\text{TECH} = \left(\frac{D_o^t(x^{t+1}, y^{t+1})}{D_o^{t+1}(x^{t+1}, y^{t+1})} \frac{D_o^t(x^t, y^t)}{D_o^{t+1}(x^t, y^t)} \right)^{1/2} \tag{5.53}$$

so that

$$M_o\left(x^t, y^t, x^{t+1}, y^{t+1}\right) = \text{EFFCH} \cdot \text{TECH}. \tag{5.54}$$

To estimate the index and its components requires solving four linear programming problems for each DMU k' for each pair of years, t and $t + 1$. To simplify, denote time by τ, so that $\tau = t$ or $t + 1$.

$$\begin{aligned} \left(D_o^t\left(x_{k'}^t, y_{k'}^t\right)\right)^{-1} &= \max \ \lambda \\ \text{s.t.} \lambda y_{k'm}^t &\leqq \sum_{k=1}^{K} z_k y_{km}^\tau, \quad m = 1, \ldots, M \\ \sum_{k=1}^{K} z_k x_{kn}^\tau &\leqq x_{k'n}^t, \quad n = 1, \ldots, N \\ z_k &\geqq 0, \quad k = 1, \ldots, K \end{aligned} \tag{5.55}$$

and

$$\begin{aligned} \left(D_o^\tau\left(x_{k'}^{t+1}, y_{k'}^{t+1}\right)\right)^{-1} &= \max \ \lambda \\ \text{s.t.} \lambda y_{k'm}^{t+1} &\leqq \sum_{k=1}^{K} z_k y_{km}^\tau, \quad m = 1, \ldots, M \\ \sum_{k=1}^{K} z_k x_{kn}^\tau &\leqq x_{k'n}^{t+1}, \quad n = 1, \ldots, N \\ z_k &\geqq 0, \quad k = 1, \ldots, K. \end{aligned} \tag{5.56}$$

When $\tau = t$ we have the estimates of the first pair of distance functions and when $\tau = t + 1$ we obtain the second pair.[13]

Before continuing with our decomposition of the Malmquist productivity change index, we look at two economic interpretations of it. We begin by relating it to the Solow "residual" and then we turn to the Törnqvist productivity index. The latter connection is one of the main results in Caves et al. (1982).

[13] There are a number of software packages to estimate the Malmquist index and DEA problems in general, see R. Barr (2004) for a discussion.

Returning to Solow, we assume as he did that the technology T can be represented by a production function, i.e., we assume that a single output is produced by multiple inputs. It is defined as

$$F(x) = \max\{y : (x, y) \in T\} \tag{5.57}$$

and given standard assumptions the maximum exists. Note also that this function can be related to the distance functions used to define the Malmquist index, i.e.,

$$\begin{aligned} F(x) &= \max\{y : D_o(x, y) \leqq 1\} \\ &= \max\left\{y : D_o(x, 1) \leqq \frac{1}{y}\right\} \\ &= 1/D_o(x, 1). \end{aligned} \tag{5.58}$$

Thus, the output distance function may be written as

$$D_o(x, y) = y/F(x). \tag{5.59}$$

In addition we assume, like Solow, that technical change is Hicks-neutral,[14] so the production function may be written as

$$y^\tau = A(\tau)F(x^\tau), \quad \tau = t \text{or } t + 1. \tag{5.60}$$

Inserting this into the definition of the Malmquist productivity index (see below), and then substituting $y^\tau/A(\tau) = F(x^\tau)$ from above yields

$$\begin{aligned} M_o\left(x^t, y^t, x^{t+1}, y^{t+1}\right) &= \left(\frac{\dfrac{y^{t+1}}{A(t)F(x^{t+1})}}{\dfrac{y^t}{A(t)F(x^t)}} \frac{\dfrac{y^{t+1}}{A(t+1)F(x^{t+1})}}{\dfrac{y^t}{A(t+1)F(x^t)}}\right)^{1/2} \\ &= A(t + 1)/A(t). \end{aligned} \tag{5.61}$$

This shows that in the single output case, the Malmquist and the Solow approaches coincide. Note that in this case there is no efficiency change, i.e.,

$$\text{EFFCH} = 1 \tag{5.62}$$

which follows from the fact that output is on the frontier by definition of the production function. In this case, productivity change is equal to technical change.

Next, following CCD we look at the relationship between the Malmquist index and the Törnqvist productivity change indexes. Here, we need to introduce output

[14] See Chambers and Färe (1994) for various notions of neutral technical change.

prices $p \in \Re_+^M$ and input prices $w \in \Re_+^N$ for time periods t and $t+1$. Then the Törnqvist productivity change index in logarithmic form is

$$\begin{aligned}&\sum_{m=1}^{M} \frac{1}{2}\left(\frac{p_m^{t+1}y_m^{t+1}}{p^{t+1}y^{t+1}} + \frac{p_m^t y_m^t}{p^t y^t}\right)(\ln y_m^{t+1} - \ln y_m^t)\\ -&\sum_{n=1}^{N} \frac{1}{2}\left(\frac{w_n^{t+1}x_n^{t+1}}{w^{t+1}x^{t+1}} - \frac{w_n^t x_n^t}{w^t x^t}\right)(\ln x_n^{t+1} - \ln x_n^t)\end{aligned} \tag{5.63}$$

To derive this index from the Malmquist productivity change index Caves et al. (1982) assumed that the distance functions are of the translog form and that the gradient vectors of the distance functions are equal to observed prices. These assumptions, in addition to conditions on the second order terms along with Diewert's (1976) quadratic lemma yield the result.

Returning to the decomposition of the index, let us start by decomposing the technical change component and then return to decomposing efficiency change.

Technical change can be decomposed into three multiplicative parts: (a) output-biased technical change (OBTC), (b) input-biased technical change (IBTC), and (c) a magnitude component (MATC),

$$\text{TECH} = \text{OBTC} \cdot \text{IBTC} \cdot \text{MATC}. \tag{5.64}$$

Specifically, these are defined as

$$\text{OBTC} = \left(\frac{D_o^t(x^{t+1}, y^{t+1})D_o^{t+1}(x^{t+1}, y^t)}{D_o^{t+1}(x^{t+1}, y^{t+1})D_o^t(x^{t+1}, y^t)}\right)^{1/2}, \tag{5.65}$$

$$\text{IBTC} = \left(\frac{D_o^{t+1}(x^t, y^t)D_o^t(x^{t+1}, y^t)}{D_o^t(x^t, y^t)D_o^{t+1}(x^{t+1}, y^t)}\right)^{1/2}, \tag{5.66}$$

$$\text{MATC} = D_o(x^t, y^t)/D_o^{t+1}(x^t, y^t). \tag{5.67}$$

Under constant returns to scale

$$D_o(x, y) = 1/D_i(y, x) \tag{5.68}$$

thus input biased technical change can be "naturally" expressed in terms of input distance functions as

$$\text{IBTC} = \left(\frac{D_i^{t+1}(y^t, x^{t+1})}{D_i^t(y^t, x^{t+1})}\right)^{1/2}. \tag{5.69}$$

Estimating biased technical change requires solving some "new" types of distance functions which assess "hypothetical" rather than observed input output

combinations. Here, we consider the estimation of one of these, $D_o^t(x^{t+1}, y^t)$, where inputs are from period $t + 1$ and outputs from period t.

For observation k' we estimate

$$\begin{aligned} D_o^t\left(x_{k'}^{t+1}, y_{k'}^t\right)^{-1} &= \max\ \lambda \\ \text{s.t.} \sum_{k=1}^{K} z_k x_{kn}^t &\leqq x_{k'n}^{t+1}, \quad n = 1, \ldots, N \\ \sum_{k=1}^{K} z_k y_{km}^t &\geqq \lambda y_{k'm}^t, \quad m = 1, \ldots, M \\ z_k &\geqq 0, \quad k = 1, \ldots, K. \end{aligned} \tag{5.70}$$

Next, we turn our attention to the decomposition of the efficiency change component of the Malmquist index. We focus on the simple decomposition of efficiency change estimated under constant returns to scale into two components: efficiency change under variable returns to scale and scale efficiency change. Scale efficiency was introduced in Färe et al. (1983), but is often attributed to Banker (1984), who focuses on what he calls the most productive scale size, i.e., the tangency between the VRS and CRS technologies which is what we call scale efficient. Our definition is

$$S_o(x, y) = D_o(x, y|\text{VRS})/D_o(x, y|\text{CRS}), \tag{5.71}$$

where $D_o(x, y|\text{CRS})$ is as defined earlier, and estimated relative to the constant returns to scale reference technology (where the intensity variables are only restricted to be nonnegative). $D_o(x, y|\text{VRS})$ is similarly defined but with the additional constraint on the intensity variables that they sum to unity and can be estimated for an observation k' as

$$\begin{aligned} D_o\left(x^{k'}, y^{k'}|VRS\right)^{-1} &= \max\ \lambda \\ \text{s.t.} \sum_{k=1}^{K} z_k x_{kn} &\leqq x_{k'n}, \ n = 1, \ldots, N \\ \sum_{k=1}^{K} z_k y_{km} &\geqq \lambda y_{k'm}, \ m = 1, \ldots, M \\ z_k &\geqq 0, \quad k = 1, \ldots, K, \ \sum_{k=1}^{K} z_k = 1. \end{aligned} \tag{5.72}$$

The decomposition of the original efficiency change component may now be written as

$$\text{EFFCH} = \frac{S_o^t(x^t, y^t)}{S_o^{t+1}(x^{t+1}, y^{t+1})} \frac{D_o^{t+1}(x^{t+1}, y^{t+1}|\text{VRS})}{D_o^t(x^t, y^t|\text{VRS})}. \tag{5.73}$$

The first ratio is the change in scale efficiency and the second indicates efficiency change measured relative to the variable returns to scale technology. We note that there have been many proposals for alternate decompositions of the Malmquist index. For a detailed discussion see Grosskopf (2003) or Färe et al. (2008).

As shown in Sect. 5.3, there are a sequence of nonradial measures that might be used in defining the Malmquist index or its components.[15] Note that excluding the directional distance function, these measures project onto the efficient subset Eff $L(y)$, and hence they cannot be used to identify what we call the congestion component in the index decomposition (see Byrnes et al. (1984) for an application of this type of decomposition in a single period case). This follows from the definition of efficiency in the Pareto-Koopmans sense.

Regarding the directional distance function and its application to productivity measurement, one can define what has been called the Luenberger productivity indicator. We call it an indicator due to its additive rather than multiplicative structure. Recalling that efficiency is indicated by a value of zero for the directional distance function, the Luenberger productivity indicator signals productivity growth with values greater than zero, and declines less than zero and is defined as

$$\begin{aligned} L = &\frac{1}{2}\left(\vec{D}_{T^{t+1}}\left(x^t, y^t; g_x, g_y\right) - \vec{D}_{T^{t+1}}\left(x^{t+1}, y^{t+1}; g_x, g_y\right)\right) \\ &+ \frac{1}{2}\left(\vec{D}_{T^t}\left(x^t, y^t; g_x, g_y\right) - \vec{D}_{T^t}\left(x^{t+1}, y^{t+1}; g_x, g_y\right)\right). \end{aligned} \tag{5.74}$$

As with the Malmquist index, the Luenberger indicator may be estimated using DEA linear programming techniques, see Sect. 5.3.[16]

Another family of productivity indexes are based on the indirect measures of efficiency. An example is the cost indirect output distance function which adds a cost constraint to the output-based DEA-type efficiency model. Specifically, such a problem may be written for observation k' as

$$\begin{aligned} \mathrm{ID}_o\left(y^{k'}, (w/c)^{k'}\right)^{-1} = \max\ & \lambda \\ \text{s.t.} \sum_{k=1}^{K} z_k x_{kn} &\leqq x_n, \quad n = 1, \ldots, N \\ \sum_{k=1}^{K} z_k y_{km} &\geqq \lambda y_{k'm}, \quad m = 1, \ldots, M \\ \sum_{n=1}^{N} w_{k'n} x_n &\leqq c_{k'} \\ z_k &\geqq 0, \quad k = 1, \ldots, K, \end{aligned} \tag{5.75}$$

[15] Nonradial Malmquist productivity indexes are discussed by Tone (2004) and have been applied to data by, e.g., Chen (2003) and Fukuyama and Weber (2001).

[16] See Färe and Grosskopf (2004) for its uses.

where the w's are input prices and c is the budget or allowed cost. This model allows input to be optimally allocated within the given budget constraint

$$\sum_{n=1}^{N} w_{k'n} x_n \leqq c_{k'}, \tag{5.76}$$

and can be substituted for the traditional output distance functions to construct a cost indirect Malmquist productivity index.

5.5 A Dynamic Malmquist Productivity Index

In this section, we return to our discussion of the dynamic model from Sect. 5.2 and use it to define a dynamic Malmquist (output) productivity index.[17] The dynamics are introduced by allowing interdependence among technologies across time. As we showed earlier, total output at say period t may be either consumed in that period (as in the static model) or used as an input in the following period, i.e., total output in period t is either final or intermediate

$$y^t = {}^f y^t + {}^i y^t. \tag{5.77}$$

If ${}^i y^t = 0$ for all periods, then the dynamic model simplifies into a sequence of standard DEA models.

The dynamic (output-oriented) Malmquist productivity index is defined – like the usual Malmquist productivity index – in terms of distance functions, in this case, however, they are dynamic distance functions. Let $\Delta_o^\tau\left(x^t, {}^f y^t\right)$ and $\Delta_o^\tau\left(x^{t+1}, {}^f y^{t+1}\right), \quad \tau = t, t+1$ be dynamic distance functions, which scale on final outputs ${}^f y$, then the dynamic Malmquist index is

$$\Omega_o\left(x^t, {}^f y^t, x^{t+1}, {}^f y^{t+1}\right) = \left[\frac{\Delta_o^t\left(x^{t+1}, {}^f y^{t+1}\right)}{\Delta_o^t\left(x^t, {}^f y^t\right)} \frac{\Delta_o^{t+1}\left(x^{t+1}, {}^f y^{t+1}\right)}{\Delta_o^{t+1}\left(x^t, {}^f y^t\right)}\right]^{1/2}. \tag{5.78}$$

As with the traditional Malmquist index, the dynamic version may be decomposed into an efficiency change and a technical change component,

$$\begin{aligned}\Omega_o\left(x^t, {}^f y^t, x^{t+1}, {}^f y^{t+1}\right) = &\left(\frac{\Delta_o^{t+1}\left(x^{t+1}, {}^f y^{t+1}\right)}{\Delta_o^t\left(x^t, {}^f y^t\right)}\right)(\text{EFFCH}) \\ &\times \left(\frac{\Delta_o^t\left(x^{t+1}, {}^f y^{t+1}\right)}{\Delta_o^{t+1}\left(x^{t+1}, {}^f y^{t+1}\right)} \frac{\Delta_o^t\left(x^t, {}^f y^t\right)}{\Delta_o^{t+1}\left(x^t, {}^f y^t\right)}\right)^{1/2}(\text{TECH}).\end{aligned} \tag{5.79}$$

[17] This section is based on Färe and Grosskopf (2010).

In addition to the decomposition above, the efficiency change component may itself be decomposed into two dynamic parts and one static part, the latter resemble the distance functions in the traditional Malmquist efficiency change component, with the addition of the intermediate input which would be treated as exogenous. We also modify these by substituting the ${}^{f}y$ terms for the y terms in the static distance functions, including ${}^{i}y$ terms as exogenous and define efficiency change as

$$\left(\frac{\Delta_o^{t+1}\left(x^{t+1},{}^{f}y^{t+1}\right)}{\Delta_o^{t}\left(x^{t},{}^{f}y^{t}\right)}\right)=\left(\frac{D_o^{t}\left(x^{t},{}^{f}y^{t}\right)}{\Delta_o^{t}\left(x^{t},{}^{f}y^{t}\right)}\right)\left(\frac{\Delta_o^{t+1}\left(x^{t+1},{}^{f}y^{t+1}\right)}{D_o^{t+1}\left(x^{t+1},{}^{f}y^{t+1}\right)}\right)\times\left(\frac{D_o^{t+1}\left(x^{t+1},{}^{f}y^{t+1}\right)}{D_o^{t}\left(x^{t},{}^{f}y^{t}\right)}\right). \tag{5.80}$$

The first term on the right-hand side is the gain due to the dynamic reallocation over the static model in period t. The second term captures the gain in period $t+1$, and the last term is the familiar static change in efficiency. The first two terms together give the net *change* in the efficiency gain between the two periods due to the advantages of the reallocation allowed in the dynamic case relative to the static case.

Finally, we show how the dynamic distance functions are estimated. Solve for observations of data k'

$$\begin{aligned}
&\max\ \lambda^1+\lambda^2+\cdots+\lambda^T \text{ s.t.}\\
&\lambda^{1f}y^1_{k'm}+{}^{i}y^1_m\leqq\sum_{k=1}^{K}z^1_k\left(y^1_{km}+{}^{i}y^0_{km}\right),\quad m=1,\ldots,M\qquad t=1\\
&\sum_{k=1}^{K}z^1_kx^1_{kn}\leqq x^1_{k'n},\quad n=1,\ldots,N\\
&z^1_k\geqq 0,\quad k=1,\ldots,K\\
&\lambda^{2f}y^2_{k'm}+{}^{i}y^2_m\leqq\sum_{k=1}^{K}z^2_k\left(y^2_{km}+{}^{i}y^1_m\right),\quad m=1,\ldots,M\qquad t=2\\
&\sum_{k=1}^{K}z^2_kx^2_k\leqq x^2_{k'}\\
&z^2_k\geqq 0,\quad k=1,\ldots,K\\
&\lambda^{Tf}y^T_{k'm}+{}^{i}y^T_m\leqq\sum_{k=1}^{K}z^T_k\left(y^T_{km}+{}^{i}y^{T-1}_m\right),\quad m=1,\ldots,M\\
&\sum_{k=1}^{K}z^T_kx^T_{kn}\leqq x^T_{k'n},\quad n=1,\ldots,N\qquad t=T\\
&z^T_k\geqq 0,\quad k=1,\ldots,K.
\end{aligned}\tag{5.81}$$

This gives us

$$\Delta_o^t\left(x_{k'}^t, {}^f y_{k'}^t\right) = 1/\lambda^{*t}, \quad \lambda^{*t}\text{optimal.} \tag{5.82}$$

The other dynamic distance functions are solved in a similar manner.

References

Banker RD. Estimating the most productive scale size using data envelopment analysis. Eur J Oper Res. 1984;17, pp. 35–44.

Barr RS. DEA software tools and technology. In: Cooper WW, Seiford LM, Zhu J, editors. Handbook on data envelopment analysis. Boston: Kluwer Academic Publishers; 2004.

Bogetoft P, Färe R, Grosskopf S, Hayes K, Taylor L. Dynamic network DEA: an Illustration. J Oper Res Soc Jpn. 2009;52(2):147–62.

Byrnes P, Färe R, Grosskopf S. Measuring productive efficiency: an application to Illinois strip mines. Manage Sci. 1984;30(6):671–81.

Caves D, Christensen L, Diewert WE. The economic theory of index numbers and the measurement of input, output, and productivity. Econometrica. 1982;50(6):1393–414.

Chambers RG, Färe R. Hicks neutrality and trade-biased Growth – a taxonomy. J Econ Theory. 1994;64:554–67.

Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res. 1978;2:429–44.

Chen Y. A non-radial Malmquist productivity index with an illustrative application to Chinese major industries. Int J Prod Econ. 2003;83:27–35.

Diewert WE. Exact and superlative index numbers. J Econom. 1976;4:115–45.

Färe R, Grosskopf S. Directional Distance Functions and Slack-Based Measures of Efficiency: Some clarifications. Eur J Oper Res. 2010;206, p. 702.

Färe R, Grosskopf S, Logan J. The relative efficiency of Illinois public utilities. Res Ener. 1983;5, pp. 349–367.

Färe R, Grosskopf S. Intertemporal production frontiers: with dynamic DEA. Boston: Kluwer Academic Publishers; 1996.

Färe R, Grosskopf S. Malmquist productivity indexes and Fisher ideal indexes. Econ J. 1992;102 (4):158–60.

Färe R, Grosskopf S, Lindgren B, Roos P. Productivity developments in Swedish hospitals: a malmquist output index approach. In: Charnes A, Cooper W, Lewin A, Seiford L, editors. Data envelopment analysis: theory, methodology and applications. Boston: Kluwer Academic Publishers; 1989, 1994. p. 253–72.

Färe R, Grosskopf S, Fukuyama H, Margaritis D. DEA and endogenous technical change, mimeo, 2009.

Färe R, Grosskopf S, Margaritis D. Efficiency and productivity: Malmquist and more. In: Fried H, Lovell CAK, Schmidt S, editors. Measurement of productive efficiency and productivity change. New York: Oxford University Press; 2008. p. 522–638.

Färe R, Grosskopf S, Roos P. Malmquist productivity indexes: a survey of theory and practice. In: Färe R, Grosskopf S, Russell RR, editors. Index numbers: essays in honour of Sten Malmquist. Boston: Kluwer Academic Publishers; 1998.

Färe R, Lovell CAK. Measuring the technical efficiency of production. J Econ Theory. 1978;19:150–62.

Färe R, Primont D. Multi-output production and duality: theory and applications. Boston: Kluwer Academic Publishers; 1995.

Farrell MJ. The measurement of productive efficiency. J Roy Stat Soc. 1957;120(3), pp. 253–281.
Fukuyama H, Weber WL. Efficiency and productivity change in non-life insurance companies in Japan. Pacific Econ Rev. 2001;6:129–46.
Karlin S. Mathematical methods and theory of games programming and economics. Reading: Addison Wesley; 1959.
Kemeny JG, Morgenstern O, Thompson GL. A Generalization of the von Neumann model of an expanding economy. Econometrica. 1956;24:115–35.
Kumar S, Russell RR. Technical change, technological catch-up, and capital deepening: relative contributions to growth and convergence. Am Econ Rev. 2002;29:527–48.
Malmquist S. Index Numbers and Indifference Surfaces. Trabajos Estatistica. 1953;4:209–42.
Margaritis D, Färe R, Grosskopf S. Productivity, convergence and policy: a Study of OECD countries and industries. J Prod Anal. 2007;28:87–105.
Ramsey FP. A mathematical theory of saving. Econ J. 1928;38:543–59.
Shephard RW. Cost and production functions. NJ: Princeton University Press; 1953.
Shephard RW. Theory of production functions. Princeton: Princeton University Press; 1970.
Shephard RW, Färe R. Dynamic theory of production correspondences. Cambridge: Oelgeschlager, Gunn and Hain Publishers; 1980.
Solow RA. Technical change and the aggregate production function. Rev Econ Stat. 1957;39:312–20.
Tone K. A slack-based measure of efficiency in Data Envelopment Analysis. Eur J Oper Res. 2001;130:498–509.
Tone K. Malmquist Productivity Index. In: Cooper WW, Seiford LM, Zhu J, editors. Handbook on data envelopment analysis. Boston: Kluwer Academic Publishers; 2004.

Chapter 6
Qualitative Data in DEA

Wade D. Cook

Abstract In many real world applications involving performance measurement, it is necessary to deal with qualitative data factors. This chapter discusses the modeling of such factors within the DEA structure.

Keywords Data envelopment analysis • Efficiency • Rank position • Ordinal data • Qualitative data

6.1 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 et al. (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

W.D. Cook (✉)
Schulich School of Business, York University, Toronto, Canada, M3J 1P3
e-mail: wcook@schulich.yorku.ca

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_6,

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 decision-making unit (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 Sect. (6.2) 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 nonquantifiable factors need to be considered. In the context of DEA, the projects represent the DMU. 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 is evaluated. Section 6.3 examines the radial projection DEA model in the context of ordinal data. Section 6.4 discusses the application of this ordinal DEA model to the two presented problems. In Sect. 6.5, various settings involving ordinal data are discussed. Conclusions and further directions are presented in Sect. 6.6.

6.2 Problem Settings Involving Ordinal Data

6.2.1 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 five-point scale).

Let us assume that for each qualitative criterion, each initiative is rated on a five-point scale, where the particular point on the scale is chosen through a consensus

Table 6.1 Ratings by criteria

	Outputs				Inputs	
Project no.	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

Table 6.2 Potential benefits

Criteria	Subcriteria 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.

on the part of executives within the organization. Table 6.1 presents an illustration of how the data might appear for ten projects, three qualitative output criteria (benefits), identified as 1, 2, and 3, and three 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 6.2 and 6.3.

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 6.3 Potential costs

Criteria	Subcriteria or interpretation
1. Technical expertise available internally	
2. Technical expertise available externally	Consultants Other research centers
3. Technology available	Equipment Software

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 five-point scale, it is generally understood that this scale merely provides a relative positioning of the projects. In a number of agencies investigated (e.g., hydroelectric and telecommunications companies), five-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.

6.2.2 *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

Table 6.4 Data for telephone offices

DMU no.	X_1	X_2	X_3	Y_1	Y_2	Y_3	Y_4	Y_5
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

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 five-point Likert scale. Table 6.4 displays the data.

In the section to follow the conventional DEA structure is adapted to accommodate variables measured on an ordinal scale.

6.3 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 DMUs, $k = 1, \ldots, N$ 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.

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 et al. (1984), and stating it in ratio form, the efficiency of DMU "o" follows from the solution of

$$
\begin{aligned}
e_o = \max\ & \left(\mu_o + \sum_{r\in R_1} \mu_r^1 y_{ro}^1 + \sum_{r\in R_2} \mu_r^2 y_{ro}^2\right) \Big/ \left(\sum_{i\in I_1} v_i^1 x_{io}^1 + \sum_{i\in I_2} v_i^2 x_{io}^2\right) \\
\text{s.t.}\ & \left(\mu_{\mathrm{o}} + \sum_{r\in R_1} \mu_r^1 y_{rk}^1 + \sum_{r\in R_2} \mu_r^2 y_{rk}^2\right) \Big/ \left(\sum_{i\in I_1} v_i^1 x_{ik}^1 + \sum_{i\in I_2} v_i^2 x_{ik}^2\right) \le 1, \ \text{all } k \\
& \mu_r^1, \mu_r^2, v_i^1, v_i^2 \ge \varepsilon, \ \text{all } r, i
\end{aligned}
\tag{6.1}
$$

where $\varepsilon > 0$ is the "non-Archimedian infinitesimal" described after (1.2) in Chap. 1.

Problem (6.1) is convertible to the linear programming format:

$$
\begin{aligned}
e_o = \max\ & \mu_o + \sum_{r\in R_1} \mu_r^1 y_{ro}^1 + \sum_{r\in R_2} \mu_r^2 y_{ro}^2 \\
\text{s.t.}\ & \sum_{i\in I_1} v_i^1 x_{io}^1 + \sum_{i\in I_2} v_i^2 x_{io}^2 = 1 \\
& \mu_o + \sum_{r\in R_1} \mu_r^1 y_{rk}^1 + \sum_{r\in R_2} \mu_r^2 y_{rk}^2 - \sum_{i\in I_1} v_i^1 x_{ik}^1 + \sum_{i\in I_2} v_i^2 x_{ik}^2 \le 0, \quad \text{all } k \\
& \mu_r^1, \mu_r^2, v_i^1, v_i^2 \ge \varepsilon, \quad \text{all } r, i,
\end{aligned}
\tag{6.2a}
$$

whose dual is given by

$$\begin{aligned}
\min \quad & \theta - \varepsilon \sum_{r \in R_1 U R_2} s_r^+ - \varepsilon \sum_{i \in I_1 U I_2} s_i^- \\
\text{s.t.} \quad & \sum_{k=1}^{N} \lambda_k y_{rk}^1 - s_r^+ = y_{ro}^1, \quad r \in R_1 \\
& \sum_{n=1}^{N} \lambda_k y_{rk}^2 - s_r^+ = y_{ro}^2, \quad r \in R_2 \\
& \theta x_{io}^1 - \sum_{k=1}^{N} \lambda_k x_{ik}^1 - s_i^- = 0, \quad i \in I_1 \\
& \theta x_{io}^2 - \sum_{k=1}^{N} \lambda_k x_{ik}^2 - s_i^- = 0, \quad i \in I_2 \\
& \sum_{k=1}^{N} \lambda_k = 1 \\
& \lambda_k, s_r^+, s_i^- \geq 0, \text{all } k, r, i \; \theta, \text{unrestricted}
\end{aligned} \tag{6.2b}$$

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 Sect. 6.2.2, "customer satisfaction," y_5 is measured on a five-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 Sect. 6.5, 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 (6.1) (and (6.2a)) 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 handling *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 (6.2a) 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

$$\gamma_{rk}(\ell) = \begin{cases} 1 & \text{if DMU } k \text{ is ranked in } \ell\text{th position on output } r \\ 0, & \text{otherwise} \end{cases}$$

$$\delta_{ik}(\ell) = \begin{cases} 1 & \text{if DMU } k \text{ is ranked in } \ell\text{th position on input } i \\ 0, & \text{otherwise.} \end{cases}$$

For example, if a five-point scale is used, and if DMU #1 is ranked in $\ell = 3$rd place on ordinal output $r = 5$, then $\gamma_{51}(3) = 1$, $\gamma_{51}(\ell) = 0$, for all other rank positions ℓ. Thus, y_{51}^2 is assigned the value $y_5^2(3)$, the *worth* to be credited to the 3rd rank position on output factor 5. It is noted that y_{rk}^2 can be represented in the form

$$y_{rk}^2 = y_r^2(\ell_{rk}) = \sum_{\ell=1}^{L} y_r^2(\ell)\gamma_{rk}(\ell),$$

where ℓ_{rk} is the rank position occupied by DMU k on output r. Hence, model (6.2a) can be rewritten in the more representative format:

$$\begin{aligned} e_o = \max\ & \mu_o + \sum_{r\in R_1} \mu_r^1 y_{ro}^1 + \sum_{r\in R_2}\sum_{\ell=1}^{L} \mu_r^2 y_r^2(\ell)\gamma_{ro}(\ell) \\ \text{s.t.}\quad & \sum_{i\in I_1} \upsilon_i^1 x_{io}^1 + \sum_{i\in I_2}\sum_{\ell=1}^{L} \upsilon_i^2 x_i^2(\ell)\delta_{io}(\ell) = 1 \\ \mu_o + \sum_{r\in R_1} \mu_r^1 y_{rk}^1 + \sum_{r\in R_2}\sum_{\ell=1}^{L} \mu_r^2 y_r^2(\ell)\gamma_{rk}(\ell) - \sum_{i\in I_1} \upsilon_i^1 x_{ik}^1 - \sum_{i\in I_2}\sum_{\ell=1}^{L} \upsilon_i^2 x_i^2(\ell)\delta_{ik}(\ell) \le 0, & \quad \text{all } k \\ \{Y_r^2 = (y_r^2(\ell)),\quad X_i^2 = (x_i^2(\ell))\} \in \Psi & \\ \mu_r^1, \upsilon_i^1 \ge \varepsilon. & \end{aligned} \tag{6.3}$$

In (6.3) 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_{i1k}^1 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.

6.3.1 Permissible Worth Vectors

The values or worths $\{y^2_r(\ell)\}$, $\{x^2_i(\ell)\}$, 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 Pth position than in the $(\ell+1)$th position on any such ordinal factor. Specifically, $y^2_r(\ell) > y^2_r(\ell+1)$ and $x^2_i(\ell) < x^2_i(\ell+1)$. That is, for outputs, one places a higher weight on being ranked in ℓth place than in $(\ell+1)$th place. For inputs, the opposite is true. A set of linear conditions that produce this realization is defined by the set Ψ, where

$$\Psi = \{(Y^2_r, X^2_r) | y^2_r(\ell) - y^2_r(\ell+1) \geq \varepsilon, \quad \ell = 1, \ldots, L-1, \quad y^2_r(L) \geq \varepsilon,$$
$$x^2_i(\ell+1) - x^2_i(\ell) \geq \varepsilon, \quad \ell = 1, \ldots, L-1, \quad x^2_i(1) \geq \varepsilon\}.$$

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 (6.3) can be written as a linear programming problem.

Theorem 6.1. Problem (6.3), in the presence of the permissible worth space Ψ, can be expressed as a linear programming problem.

Proof. In (6.3), make the change of variables

$$w^1_{r\ell} = \mu^2_r y^2_r(\ell), \qquad w^2_{i\ell} = \upsilon^2_i x^2_i(\ell).$$

It is noted that in Ψ, the expressions

$$y^2_r(\ell) - y^2_r(\ell+1) \geq \varepsilon, \quad y^2_r(L) \geq \varepsilon$$

can be replaced by

$$\mu^2_r y^2_r(\ell) - \mu^2_r y^2_r(\ell+1) \geq \mu^2_r \varepsilon, \quad \mu^2_r y^2_r(L) \geq \mu^2_r \varepsilon,$$

which becomes

$$w^1_{r\ell} - w^1_{r\ell+1} \geq \mu^2_r \varepsilon, \quad w^2_{rL} \geq \mu^2_r \varepsilon.$$

A similar conversion holds for the $x^2_i(\ell)$.

Problem (6.3) now becomes

$$\begin{gathered}
e_o = \max \mu_o + \sum_{r\in R_1} \mu^1_r y^1_{ro} + \sum_{r\in R_2} \sum_{\ell=1}^{L} w^1_{r\ell} \gamma_{ro}(\ell) \\
\text{s.t.} \sum_{i\in I_1} v^1_i x^1_{io} + \sum_{i\in I_2} \sum_{\ell=1}^{L} w^2_{i\ell} \delta_{io}(\ell) = 1 \\
\mu_o + \sum_{r\in R_1} \mu^1_r y^1_{rk} + \sum_{r\in R_2} \sum_{\ell=1}^{L} w^1_{rl} \gamma_{rk}(\ell) - \sum_{i\in I_1} v^1_i x^1_{ik} - \sum_{i\in I_2} \sum_{\ell=1}^{L} w^2_{i\ell} \delta_{ik}(\ell) \leq 0, \quad \text{all } k \\
w^1_{r\ell} - w^1_{r\ell+1} \geq \mu^2_r \varepsilon, \quad \ell = 1, \ldots, L-1, \quad \text{all } r \in R_2 \\
w^1_{rL} \geq \mu^2_r \varepsilon, \quad \text{all } r \in R_2 \\
w^2_{i\ell+1} - w^2_{i\ell} \geq v^2_i \varepsilon, \quad \ell = 1, \ldots, L-1, \quad \text{all } i \in I_2 \\
w^2_{i1} \geq v^2_i \varepsilon, \quad \text{all } i \in I_2 \\
\mu^1_r, v^1_i \geq \varepsilon, \quad \text{all } r \in R_1, i \in I_1 \\
\mu^2_r, v^2_i \geq \varepsilon, \quad \text{all } r \in R_2, i \in I_2. \qquad (6.4)
\end{gathered}$$

Problem (6.4) is clearly in linear programming problem format.

We state without proof the following theorem.

Theorem 6.2. At the optimal solution to (6.4), $\mu^2_r = v^2_i = \varepsilon$ for all $r \in R_2$, $i \in I_2$.

Problem (6.4) can then be expressed in the form:

$$\begin{gathered}
e_o = \max \mu_o + \sum_{r\in R_1} \mu^1_r y^1_{ro} + \sum_{r\in R_2} \sum_{\ell=1}^{L} w^1_{r\ell} \gamma_{ro}(\ell) \\
\text{s.t.} \sum_{i\in I_1} v^1_i x^1_{io} + \sum_{i\in I_2} \sum_{\ell=1}^{L} w^2_{i\ell} \delta_{io}(\ell) = 1 \\
\mu_o + \sum_{r\in R_1} \mu^1_r y^1_{rk} + \sum_{r\in R_2} \sum_{\ell=1}^{L} w^1_{r\ell} \gamma_{rk}(\ell) - \\
\sum_{i\in I_1} v^1_i x^1_{ik} - \sum_{i\in I_2} \sum_{\ell=1}^{L} w^2_{i\ell} \delta_{ik}(\ell) \leq 0, \quad \text{all } k \\
-w^1_{r\ell} + w^1_{r\ell+1} \leq -\varepsilon^2, \quad \ell = 1, \ldots, L-1, \quad \text{all } r \in R_2 \\
-w^1_{rL} \leq -\varepsilon^2, \quad \text{all } r \in R_2 \\
-w^2_{i\ell+1} + w^2_{i\ell} \leq -\varepsilon^2, \quad \ell = 1, \ldots, L-1, \quad \text{all } i \in I_2 \\
-w^2_{i1} \leq -\varepsilon^2, \quad \text{all } i \in I_2 \\
\mu^1_r, v^1_i \geq \varepsilon, \quad r \in R_1, \quad i \in I_1. \qquad (6.5a)
\end{gathered}$$

It can be shown that (6.5a) is equivalent to the *standard* VRS model. First, we form the dual of (6.5a).

$$
\begin{aligned}
\min\ \theta - \varepsilon \sum_{r\in R_1} s_r^+ - \varepsilon \sum_{i\in I_1} s_i^- - \varepsilon^2 \sum_{r\in R_2}\sum_{\ell=1}^{L} \alpha_{r\ell}^1 - \varepsilon^2 \sum_{i\in I_2}\sum_{\ell=1}^{L} \alpha_{i\ell}^2 \\
\text{s.t.} \sum_{k=1}^{N} \lambda_k y_{rk}^1 - s_r^+ = y_{ro}^1, \quad r \in R_1 \\
\theta x_{io}^1 - \sum_{k=1}^{N} \lambda_k x_{ik}^1 - s_i^- = 0, \quad i \in I_1 \\
\left.\begin{aligned}
&\sum_{k=1}^{N} \lambda_k \gamma_{rk}(1) - \alpha_{r1}^1 = \gamma_{ro}(1) \\
&\sum_{k=1}^{N} \lambda_k \gamma_{rk}(2) + \alpha_{r1}^1 - \alpha_{r2}^1 = \gamma_{ro}(2) \\
&\qquad\vdots \\
&\sum_{k=1}^{N} \lambda_k \gamma_{rk}(L) + \alpha_{rL-1}^1 - \alpha^1{}_{rL} = \gamma_{ro}(L)
\end{aligned}\right\} r \in R_2 \\
\left.\begin{aligned}
&\delta_{io}(L)\theta - \sum_{k=1}^{N} \lambda_k \delta_{ik}(L) - \alpha_{iL}^2 = 0 \\
&\delta_{io}(L-1)\theta - \sum_{k=1}^{N} \lambda_k \delta_{ik}(L-1) + \alpha_{iL}^2 - \alpha_{iL-1}^2 = 0 \\
&\qquad\vdots \\
&\delta_{io}(1)\theta - \sum_{k=1}^{N} \lambda_k \delta_{ik}(1) + \alpha_{i2}^2 - \alpha_{i1}^2 = 0
\end{aligned}\right\} i \in I_2 \\
\sum_{k=1}^{N} \lambda_k = 1 \\
\lambda_k,\ \mathrm{s}_r^+,\ \mathrm{s}_i^-, \alpha_{r\ell}^1, \alpha_{i\ell}^2 \geq 0 \\
\theta \text{ unrestricted.}
\end{aligned}
\tag{6.5b}
$$

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 (6.5b) by replacing the ℓth 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

$$\overline{\gamma}_{rk}(\ell) = \sum_{n=1}^{\ell} \gamma_{rk}(n) = \gamma_{rk}(1) + \gamma_{rk}(2) + \cdots + \gamma_{rk}(\ell),$$

and

$$\overline{\delta}_{ik}(\ell) = \sum_{n=\ell}^{L} \delta_{ik}(n) = \delta_{ik}(L) + \delta_{ik}(L-1) + \cdots + \delta_{ik}(\ell),$$

problem (6.5b) can be rewritten as:

$$\begin{aligned}
\min \theta - \varepsilon \sum_{r\in R_1} s_r^+ - \varepsilon \sum_{i\in I_1} s_i^- - \varepsilon^2 \sum_{r\in R_2} \sum_{\ell=1} \alpha_{r\ell}^1 - \varepsilon^2 \sum_{i\in I_2} \sum_{\ell=1} \alpha_{i\ell}^2 \\
\text{s.t.} \sum_{k=1}^{N} \lambda_k y_{rk}^1 - s_r^+ = y_{ro}^1, \quad r \in R_1 \\
\theta x_{io}^1 - \sum_{k=1}^{N} \lambda_k x_{ik}^1 - s_i^- = 0, \quad i \in I_1 \\
\sum_{k=1}^{N} \lambda_k \overline{\gamma}_{rk}(\ell) - \alpha_{r\ell}^1 = \overline{\gamma}_{ro}(\ell), \quad r \in R_2, \quad \ell = 1, \ldots, L \\
\theta \overline{\delta}_{io}(\ell) - \sum_{k=1}^{N} \lambda_k \overline{\delta}_{ik}(\ell) - \alpha_{i\ell}^2 = 0, \quad i \in I_2, \quad \ell = 1, \ldots, L \\
\sum_{k=1}^{N} \lambda_k = 0 \\
\lambda_k, s_r^+, s_i^-, \alpha_{r\ell}^1, \alpha_{i\ell}^2 \geq 0, \quad \text{all } i, r, \ell, k, \theta \text{ unrestricted in sign.}
\end{aligned} \tag{6.6a}$$

The dual of (6.6a) has the VRS format:

$$\begin{aligned}
e_o = \max \mu_o + \sum_{r\in R_1} \mu_r^1 y_{ro}^1 + \sum_{r\in R_2} \sum_{\ell=1}^{L} w_{r\ell}^1 \overline{\gamma}_{ro}(\ell) \\
\text{s.t.} \quad \sum_{i\in I_1} \upsilon_i^1 x_{io}^1 + \sum_{i\in I_2} \sum_{\ell=1}^{L} \bar{w}_{r\ell}^2 \overline{\delta}_{io}(\ell) = 1 \\
\mu_o + \sum_{r\in R_1} \mu_r^1 y_{rk} + \sum_{r\in R_2} \sum_{\ell=1}^{L} w_{r\ell}^1 \overline{\gamma}_{rk}(\ell) - \sum_{i\in I_1} \upsilon_i^1 x_{ik}^1 - \sum_{i\in I_2} \sum_{\ell=1}^{L} w_{i\ell}^2 \overline{\delta}_{ik}(\ell) \leq 0, \quad \text{all } k \\
\mu_r^1, \upsilon_i^1 \geq \varepsilon, w_{r\ell}^1, w_{i\ell}^2 \geq \varepsilon^2,
\end{aligned} \tag{6.6b}$$

which is a form of the VRS model. The slight difference between (6.6b) 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 (6.6a) 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 (6.6b). Of course, since the intention is to have an infinitesimal lower bound on multipliers (i.e., $\varepsilon>0$), 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 (6.6b) where all multipliers have the same infinitesimal lower bounds, making (6.6b) precisely a VRS model in the spirit of Banker et al. (1984).

It is interesting to note that the IDEA approach of Cooper et al (1999) essentially involves tackling problem (6.2a) 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 (6.2a) 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 the same final point. Cook and Zhu (2006) examine both continuous and discrete projections. They show that in the case of continuous projections, the IDEA methodology is equivalent to that put forward by Cook et al (1993, 1996).

6.3.2 *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_r^2 by $\mu^2(m)$, and υ_i^2 by $\upsilon^2(n)$, and restrict:

$$\mu^2(m) - \mu^2(m+1) \geq \varepsilon, \quad m = 1, \ldots, L_1 - 1$$
$$\mu^2(L_1) \geq \varepsilon$$

and

$$\upsilon^2(n) - \upsilon^2(n+1) \geq \varepsilon, \quad n = 1, \ldots, L_2 - 1$$
$$\upsilon^2(L_2) > \varepsilon.$$

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

$$w_{r\ell}^1 = \mu^2(m_r) y_r^2(\ell),$$

$$w_{i\ell}^2 = \upsilon^2(n_i) x_i^2(\ell).$$

The corresponding version of model (6.4) would see the lower bound restrictions $\mu_r^2, \upsilon_i^2 \geq \varepsilon$ replaced by the above constraints on $\mu^2(m)$ and $\upsilon^2(n)$. Again, arguing that at the optimum in (6.4), these variables will be forced to their lowest levels, the resulting values of the $\mu^2(m)$, $\upsilon^2(n)$ will be

$$\mu^2(m) = (L_1 + 1 - m)\varepsilon, \qquad \upsilon^2(n) = (L_2 + 1 - n)\varepsilon.$$

This implies that the lower bound restrictions on $w_{r\ell}^1$, $w_{i\ell}^2$ become

$$w_{r\ell}^1 \geq (L_1 + 1 - m_r)\varepsilon^2, \qquad w_{i\ell}^2 \geq (L_2 + 1 - n_i)\varepsilon^2.$$

We now apply the above concepts to the data for the two problem settings discussed earlier.

6.4 Solutions to Applications

6.4.1 R&D Project Efficiency Evaluation

When model (6.6a) is applied to the data of Table 6.1, the efficiency scores obtained are as shown in Table 6.5.

Table 6.5 Efficiency scores (nonranked 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

Table 6.6 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

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 6.5. When a very small value of ε ($\varepsilon = 0.001$) 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 6.1 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 6.6.

Hence, criteria importance restrictions can have an impact on the efficiency status of the projects.

6.4.2 Evaluation of Telephone Office Efficiency

The data of Table 6.4 has been evaluated using Model (6.6a). Both CRS and VRS models were applied, the results of which are presented in Table 6.7.

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 6.7 Efficiency scores

DMU#	CRS score	VRS score	VRS score-constrained
1	1	1	1
2	1	1	1
3	1	1	1
4	0.927	1	0.973
5	1	1	0.921
6	0.907	0.994	0.906
7	0.848	0.849	0.823
8	0.668	0.670	0.644
9	0.848	0.970	0.885
10	0.617	0.747	0.731
11	0.763	0.815	0.716
12	1	1	0.915
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	0.886
17	0.898	1	1
18	0.928	1	0.935
19	0.993	0.993	0.961
20	1	1	1
21	1	1	1
22	1	1	1
23	0.846	1	1
24	0.918	1	0.904
25	1	1	1
26	1	1	0.955
27	0.824	0.937	0.926
28	0.954	1	0.919
29	0.949	1	1
30	1	1	1
31	1	1	0.907
32	1	1	1
33	0.962	1	1

6.5 Problem Settings and Issues Involving Qualitative Data

Qualitative data arises in a multitude of problem settings. As well, some of these practical settings can involve complex issues that are not immediately compatible with the standard model as presented above. Thus, handling the realities of ordinal data in a practical setting often does not automatically simply entail applying the models of Sect. 6.3. In this section, we examine some case examples of situations involving qualitative data, and explore some of the aforementioned complexities.

6.5.1 Implementation of Robotics: Identifying Efficient Implementers

Cook et al. (1992) examine robotics implementations in 30 companies, seeking to determine which among these are the most efficient implementers. The study is based upon three types of variables: (1) inputs or initial conditions prior to installation; (2) outputs or outcomes of the end of the installation; and (3) environmental factors that were treated here as control variables *after* the DEA analysis was completed.

The initial conditions at the start of the project (*inputs*) were

- System complexity: A count of four components of robotics systems.
- Previous experience with technology: A summed score of 4 five-point Likert scale measures.
- Novelty of the application: A Likert scale measure of the installation's innovativeness.

The *outputs* in this study are the three outcomes at the end of the project's installation, namely:

- Startup time: Number of weeks required to take the technology from physical installation to routine use.
- Uptime: Percent of total production time, facility is available.
- Management satisfaction: A perceptual measure on a five-point Likert scale.

A major difficulty in applying DEA in this setting was developing a rationalization of multiple factors, with some being quantitative and some having Likert scale measures. In the case of the input "Previous Experience," which contained *multiple* Likert scale parameters, the sum of the scale values was finally treated as a quantitative variable.

Part of this case study, as well, was the analysis of efficiency in terms of various *control variables*, namely, supplier management, plant size, and urgency. By examining the average ratings for a subgroup of installations when the DEA model was run only for that subset (e.g., those in small plants), versus the average for that same group as part of the total set of 30 plants, one is able to draw conclusions about the most effective settings in which to undertake robotics installations.

6.5.2 A Fair Model for Aggregating Preferential Votes

Cook and Kress (1990) examine the use of DEA to prioritize candidates, (the DMUs), in a preferential election. Here, each voter is asked to select a subset R of K candidates, and to rank order these candidates from most to least preferred. By its very nature, the data in this setting is of the ordinal or rank order type. Such a voting format is common in municipal elections where a number of candidates are required to fill various positions.

In this setting, each candidate receives some number υ_{1k} of first place votes, υ_{2k} of second place votes, . . ., υ_{Rk} of Rth place votes. The objective is to derive a score e_k for each candidate k which rationally accounts for the numbers of 1st, 2nd, . . ., Rth place votes received by that candidate. The complexity here was that each voter provided a *partial ranking* only of the candidates. Unlike the previous application on robotics, where each DMU is evaluated in terms of several factors (inputs and outputs), in the current application, the voters may be thought of as the evaluation factors. Unfortunately, in most voting situations the number of such factors (voters) is very large. Thus, from a practical standpoint, the model of Sect. 6.3 cannot be applied directly. The approach taken here was to *total* the number of "hits" received by a candidate at each ordinal rank position (1, 2, . . ., R). In the context of the usual DEA structure, this is equivalent to saying that all voters get the same weight.

The model solved in Cook and Kress (1990) for deriving e_k was

$$
\begin{aligned}
e_o = \max & \sum_{r=1}^{R} w_r \upsilon_{ro} \\
\text{s.t.} & \sum_{r=1}^{R} w_r \upsilon_{rk} \leq 1, \quad \forall k \\
w_r - w_{r+1} & \geq d(r, \varepsilon), \quad r = 1, \ldots, R-1 \\
w_r & \geq d(r, \varepsilon),
\end{aligned}
$$

where $d(r, \varepsilon)$, the *discrimination intensity function*, is intended to reflect the minimum allowable gap between the worths associated with the rth and $r + 1$th rank positions. The above constraints involving $d(r, \varepsilon)$ constitute (in this application) the set Ψ as discussed in Sect. 6.3.

6.5.3 Multiple Criteria Decision Modeling: Ordinal Data, Criteria Importance, and Criteria Clearness

Certain types of DEA problems may involve *only outputs* (no inputs). The previous application in ranking candidates in a preferential voting situation is one such example. A special class of such problems is the set of multiple criteria decision problems. In multiple criteria decision settings, the decision maker is generally required to evaluate each member of a set K of alternatives in terms of various criteria. The objective is normally to create a final overall ranking of the alternatives, or at least to pick a winner. In cases where ordinal data are present, at least two issues arise. One of these issues, as addressed in the previous example on preferential voting, involves incorporating known information on the relative importance of the criteria. The AR (assurance region) restrictions in Ψ, as discussed in Sect. 1.4.3, are designed to capture such information. In many settings,

an additional feature involving the criteria may be present, namely, a ranking of those criteria in terms of the *degree of clearness* whereby the alternatives can be evaluated in terms of such criteria. That is, for a criterion that is very *fuzzy*, the resulting ranking of the alternatives (in terms of that criterion) would be less reliable, hence less important than would be true of a ranking in terms of a more clear (less fuzzy) criterion.

Cook and Kress (1991) (and later (1994)) present a multiple criteria model for ranking alternatives when ordinal data are present, and where criteria can be ordinally ranked, both in terms of importance and clearness. Their paper models the ranking problem utilizing the DEA philosophy of treating each alternative as a DMU, and maximizes the overall score of each alternative. AR constraints are specifically imposed that take criteria importance and criteria clearness into account, hence enlarging the worth vector restriction set as reflected in Ψ.

The following are some case examples where ordinal data arises naturally, and where both criteria importance and clearness are an issue.

6.5.3.1 Evaluating Vendors for Complex Systems

Cook and Johnston (1992) examine the prioritization of a set of six vendors who have submitted bids on the development of a piece of automatic testing equipment (ATE) that was to be designed for Nortel Incorporated (then called Northern Telecom). Forty different criteria were being used to evaluate the proposals; these were divided into three classes – vendor issues, ATE specifications, and delivery factors.

Historically, Nortel had, in such situations, scored vendors on a Likert scale, and rated criteria on a ten-point scale. A weighted score would then be computed for each alternative (e.g., vendor) under consideration. In the particular vendor application under discussion, the company wanted to approach the vendor choice issue in a somewhat more scientific way, given that certain criteria (e.g., vendor reliability) were less "clear," hence less reliable, than was true of other factors such as experience in designing ATE products. As a result, the aforementioned structure with the enlarged AR set Ψ, and treating the vendors as the DMUs, was applied.

6.5.3.2 Country Risk Evaluation

Cook and Hebner (1993a, b) examine the use of this same model structure to evaluate and prioritize 100 countries (the DMUs) in terms of their risk level relating to investment. The data were made available by the Japan Bond Research Institute. Fourteen criteria were used that reflect a country's risk, including (1) social stability, (2) political stability, (3) fiscal policy, etc.

The actual data available on countries' risk positions for any criterion were given in the form of percentages (e.g., one of the criteria was the growth

potential expressed as the expected rate of real per capita GNP growth in the coming year). To accomplish an overall ranking of countries in this setting, the percentage data for each criterion was translated first to an ordinal ranking of the countries. Then, the actual numerical data were used to generate criteria importance discrimination factors to allow for appropriate gaps between rank positions. Again, the decision maker was asked to rank criteria both in terms of importance and clearness, hence creating the appropriate AR restrictions defining Ψ.

6.5.3.3 Mutual Fund Selection

Cook and Hebner (1993a, b) developed a multiple criteria selection model for prioritizing a set of mutual funds, treating the funds as DMUs and the selection criteria as the outputs. The criteria include both quantitative indicators such as front end loading fees and the standard deviation of a fund's rate of return, as well as qualitative factors such as the service quality of the fund manager, in terms of his/her understanding of financial markets. The approach here was similar to that used in the country risk setting. Specifically, output data were represented by the rank position occupied by each fund on each criterion. Available cardinal data for those criteria such as front end fees, were used to generate criteria discrimination parameters, leading to assurance region restrictions.

6.5.3.4 Ordinal Data in Multicriteria Modeling: Evaluation in Terms of Subsets of Criteria

Returning again to the special class of DEA models involving multiple criteria evaluation of a set of alternatives, it is often the case that for any given DMU (alternative), only a proper subset of criteria may be relevant to that evaluation. Consider the example of the ranking of a set of R&D projects in a *pure output* setting (no inputs involved). Some projects may, for example, involve the development of nuclear technology while others may not. If a given criterion relates to environmental impacts, such as risk of a nuclear accident, such a criterion would be relevant only for projects concerning nuclear aspects. Specifically, a project that had no such involvement must be rated on only a *subset* of the criteria.

Cook et al. (1997) examine the modeling of DEA problems in the presence of ordinal factors when DMUs may be rated only on a proper subset of the outputs (criteria). The standard model of Sect. 6.3 is not immediately applicable, and the suggested approach uses the concept of the *ideal* rank position (w_{r1}) on each criterion r applicable to the DMU in question. That is, rank position "1" is the ideal or best possible ranking achievable.

Specifically, let $R_k \subseteq R$, denote the set of outputs or criteria against which DMU k will be evaluated. Then, solve the ratio problem:

$$e_o = \max e_o = \max \sum_{r \in R_o} \sum_{\ell=1}^{L} w_{r\ell} \gamma_{ro}(\ell) / \sum_{r \in R_o} w_{r1}$$

$$\text{s.t.} \sum_{r \in R_k} \sum_{\ell=1}^{L} w_{r\ell} \gamma_{rk}(\ell) / \sum_{r \in R_k} w_{r1} \leq 1, \quad \forall k$$

$$w_{r\ell} \in \Psi.$$

This problem is translatable into a linear programming framework in the usual manner.

For ordinal data problems, particularly those of the *pure output* type, this concept of measuring efficiency of a DMU relative to the ideal position of that DMU, provides an innovative way of tackling problems where the DMU has "missing" data.

6.6 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 (6.6a) 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 $(1 - \theta) \times 100\%$ 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 (6.6a) 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

Ali AI, Seiford L. Computational accuracy and infinitesimals in data envelopment analysis. Working paper. University of Massachusetts at Amherst, MA; 1993.

Banker RD, Charnes A, Cooper WW. Some models for technical and scale efficiencies in data envelopment analysis. Manage Sci. 1984;30:1078–92.

Cook WD, Doyle J, Green R, Kress M. Eur J Oper Res. 1997;93(3):602–9.
Cook WD, Hababou M. Sales performance measurement in bank branches. OMEGA Int J Manage Sci. 2001;29:299–307.
Cook WD, Hababou M, Tuenter H. Multicomponent efficiency measurement and shared inputs in data envelopment analysis: an application to sales and service performance in bank branches. J Prod Anal. 2000;14(3):2000.
Cook WD, Hebner K. A multicriteria approach to country risk evaluation: With an example employing Japanese data. Int Rev Econ Finan. 1993a;2(4):327–48.
Cook WD, Hebner K. A multicriteria approach to mutual fund selection. Finan Serv Rev. 1993b;2(1):1–20.
Cook WD, Johnston DA, McCutcheon D. Implementation of robotics: identifying efficient implementers. OMEGA Int J Manage Sci. 1992;20:227–39.
Cook WD, Johnston DA. Evaluating suppliers of complex systems: A multiple criteria approach. J Oper Res Soc. 1992;43(11):1055–61.
Cook WD, Kress M. A data envelopment model for aggregating preference rankings. Manage Sci. 1990;36(11):1302–10.
Cook WD, Kress M. A multiple criteria decision model with ordinal preference data. Eur J Oper Res. 1991;54(2):191–8.
Cook WD, Kress M. A multiple criteria composite index model for quantitative and qualitative data. Eur J Oper Res. 1994;78(3):367–79.
Cook WD, Kress M, Seiford L. On the use of ordinal data in Data Envelopment Analysis. J Oper Res Soc. 1993;44(2):133–40.
Cook WD, Kress M, Seiford L. Data Envelopment Analysis in the presence of both quantitative and qualitative factors. J Oper Res Soc. 1996;47:945–53.
Cook WD, Zhu J. Rank order data in DEA : a general framework. Eur J Oper Res. 2006;174:1021–38.
Cooper WW, Park KS, Yu G. IDEA and AR-IDEA: models for dealing with imprecise data in DEA . Manage Sci. 1999;45(4):597–607.
Cooper WW, Park KS, Yu G. IDEA (Imprecise Data Envelopment Analysis) with CMDs (column maximum decision making units). J Oper Res Soc. 2001;52:176–81.
Kim SH, Park CG, Park KS. An application of data envelopment analysis in telephone offices evaluation with partial data. Comput Oper Res. 1999;26:59–72.
Thrall RM. Duality, classification and slacks in DEA . Ann Oper Res. 1996;66:109–38.
Zhu J. Imprecise data envelopment analysis (IDEA): A review and improvement with an application. Eur J Oper Res. 2003;144:513–29.

Chapter 7
Congestion: Its Identification and Management with DEA

William W. Cooper, Honghui Deng, Lawrence M. Seiford, and Joe Zhu

Abstract Congestion is a term that is applicable in a variety of disciplines which range from medical science to traffic engineering. It has also many uses in practical everyday life. This brings with it a certain looseness in usage. We therefore expand (and refine) our discussion of congestion with reference to its use in economics where we have access to a precise meaning which we can develop in this chapter. This chapter covers the standard approaches used for treating congestion in data envelopment analysis.

Keywords Data envelopment analysis • Efficiency • Performance • Congestion

7.1 Congestion

Congestion is a term that is applicable in a variety of disciplines which range from medical science to traffic engineering. It has also many uses in practical everyday life. This brings with it certain looseness in usage. We therefore expand (and refine) our discussion of congestion with reference to its use in economics where we have access to a precise meaning which we can develop as follows.

Figure 7.1 will help us to see what is involved. Drawing on classical production theory from economics, this figure is intended to portray what is involved in a relatively simple way by restricting attention to the case of one output and one input. The horizontal axis represents the input amounts, x, and the vertical axis represents the output amounts, y. The curve portrayed in this figure relates these input amounts to the corresponding output amounts which they can produce. It is referred to as a "production function."

J. Zhu (✉)
School of Business, Worcester Polytechnic Institute, Worcester, MA 01609, USA
e-mail: jzhu@wpi.edu

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_7,

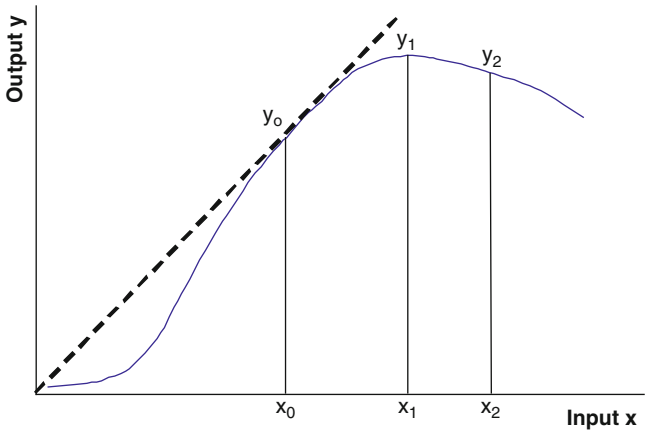


Fig. 7.1 Example of a production function

The concept of a production function in economics as graphed in Fig. 7.1 has a special meaning. It is used to relate input to output in a manner that *maximizes* the output for each input amount that may be used. See the text by Varian (1984). The classic testament can be found in Chap. IV of Samuelson (1947). This implies that no activity will occur below the "frontier" defined by this function. It is also assumed that the production surface is in the form of a smooth curve with a single peak like the one in Fig. 7.1. Empirical applications generally use models such as Cobb–Douglas functions – a log linear function – which are ever increasing in all inputs and hence have no peaks. Hence, frontier points which are technically inefficient in the sense of Chap. 1 of this handbook are assumed not to occur. (See the discussion of the horizontal segment in Figs. 1–4). Congestion, however, can occur as represented by points on the curve to the right of the peak in Fig. 7.1.

Notice that production continues to be maximal even on the declining portion of this curve. That is, at each input value beyond x_1, any point on the solid curve continues to depict the maximum output amount that can be attained. This leads to the following definition.

Definition 7.1. *Congestion* is said to occur when the output that is *maximally* possible can be increased by reducing one or more inputs without improving any other input or output. Conversely, congestion is said to occur when some of the outputs that are maximally possible are reduced by increasing one or more inputs without improving any other input or output. In Fig. 7.1, for example, this would mean that the loss in output is to be identified with the difference between y_1 and y_2 while the congesting amount of input is $x_2 - x_1$.

For a concrete example we may think of coal miners operating in an underground mine which we can relate to Fig. 7.1 as follows. Starting with input (in the form of number of miners) at $x = 0$ the output, y, measured in tons of coal, can be increased at an increasing rate until x_o is reached at output y_o. This can occur, for instance, because an increase in the number of miners makes it possible to form

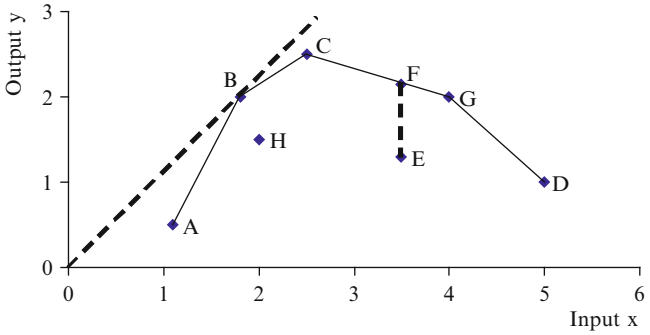


Fig. 7.2 A DEA production surface

"teams" to perform tasks in a manner that would not be possible with a smaller number of miners. From x_0 to x_1, however, total output continues to increase, but at a decreasing rate, until the maximum possible output is reached at y_1. Using more input results in a decrease from this maximum so that at x_2 we have $y_2 < y_1$ and $y_1 - y_2$ is the amount of output lost due to "congestion."

It is to be noted that these differences are all measured relative to the frontier. However, as we shall see later, it is possible to extend these concepts to identify such phenomena even when additional inefficiencies result in observations that lie below the frontiers associated with production functions.

In contrast to a smooth (everywhere differentiable) curve like the one in Fig. 7.1, the production surfaces used in data envelopment analysis (DEA) are piecewise linear (and continuous). Figure 7.2 provides an example which we can compare with Fig. 7.1 as follows. The single smooth curve exhibited in Fig. 7.1 is replaced by a series of piecewise linear segments to form a continuous boundary derived from the data in the manners that were earlier described in Chap. 1 of this handbook.

Being piecewise linear the surface in Fig. 7.2 fails to be everywhere differentiable since, in particular, derivatives fail to exist at the points where the segments are joined. Hence, we need to replace the concept of a "derivative" and the associated concept of a tangent with the more general concept of a "supporting hyperplane." An example of such a hyperplane (here a line) is exhibited by the broken line from the origin which touches the production surface at B in a manner analogous to the tangent plane (here a line) that represented the point of maximal returns to scale at y_o in Fig. 7.1. Hence, we can, if we wish, preserve the concept of "returns to scale" – which, like "congestion," is also a frontier concept in that it is associated with the frontiers of the set of production possibilities. See Chap. 2 in this handbook.

In Fig. 7.2, we refer to the "production surfaces" as "production boundaries" or "production frontiers" in place of the "production function" concept which we employed for Fig. 7.1. This is because in place of a single production function for a particular firm we can regard the production boundaries in Fig. 7.2 as a surface

generated from a series of supporting hyperplanes that touch the surface of at least one production function in a collection of production functions – which functions may differ for each of the different firms (=DMUs) represented in the data.

In DEA we do not require all points to lie on the surface, as is the case in economics. Instead, we allow points like H and E to occur, which lie below the surface empirically, as in Fig. 7.2. This means that H and E fall short of the output they could have produced with the inputs they used. Nevertheless, we shall continue to use the economic concept of congestion as in Definition 7.1. This will help us to make distinctions between "congestion" and other types of inefficiency such as "technical inefficiency," etc., in our development.

The points A, B, and C and the segments connecting them in Fig. 7.2 represent the efficiency frontier. In the approaches we use, it is this portion of the frontier which is used to evaluate the efficiency of all of the other points in the production possibility set (PPS). The segments connecting CFGD represent the congested part of the frontier. Notice that on the segments connecting A, B, and C, which together represent the efficiency frontier, it is not possible to increase (i.e., improve) the output at any point without increasing (i.e., worsening) the input. The congestion frontier has the opposite property: it is not possible to increase (i.e., improve) the output on this frontier without also decreasing (i.e., improving) the input.

The points H and E help us to understand some of the properties that will be of interest. Both points lie below the frontier. Projecting H vertically until it reaches the efficiency frontier on the segment connecting B and C shows that the point H is technically inefficient because, as evidence from the data shows, it is possible to increase the output of H without increasing its input. Alternatively, a horizontal projection from H to the segment of the efficiency frontier between A and B also shows that H is "technically inefficient" because it is possible to decrease the input without decreasing the output. These vertical and horizontal projections reflect results from "output oriented" and "input oriented" objectives respectively in the models we shall be using.

Now consider E. A horizontal projection to the segment of the frontier between A and B shows that it is possible to decrease the input at E without worsening the output. This projection, as previously noted, reflects an "input oriented" objective. It does not permit us to make distinctions like those we want to make between "technical inefficiency" and "congestion." We therefore turn to an "output oriented" objective which effects the projection from E to F. This shows that it is possible to increase the output for E without changing its input. However, this does not end the matter since we can further increase (i.e., improve) the output by decreasing (i.e., improving) the input amount for F. Indeed this increase in output along the frontier can be continued until the maximal possible output is reached, as occurs when the input is reduced to a point directly below C.

The difference between the input at E and the input at C represents the amount of congestion in the input amount used by E, while the difference between the output coordinate at C and the output coordinate at E represents the output lost due not only to congestion but also to the way it was managed. We will later use the

difference between the ordinates at C and F as our measure of the output loss due to congestion and the difference between the ordinates on E and F as the measure of output loss due to the way this input was managed. This will not only keep us in touch with economic theory, but it will also allow us to deal with problems that are of practical managerial interest. We may note, for instance, that identifying C in this (frontier) manner also produces a byproduct in the form of an estimate of the capacity that is available with fully efficient performance by the entity associated with E.

We do not pursue this topic further at this point other than to present the following as an example of managerial inefficiency in managing a congesting input.

Example. Excess inventory cluttering a factory floor in a way that interferes with production. By simply reconfiguring this excess inventory it may be possible to increase output without reducing inventory. This improvement represents the elimination of inefficiency that is caused by the way excess inventory is managed. It is the kind of inefficiency which can be represented by the movement from E to F in Fig. 7.2, and which we will identify as "managerial inefficiency." The movement from F to C will then give us a measure of the output lost due to congestion. In either case the amount of the congesting input will be measured by the difference in the value of x for E and the value of x for C.

Having now clarified what we mean by "congestion" we should probably also illustrate some of the respects in which a treatment of this topic might be of interest. A case in point is provided in the doctoral thesis of Deng (2003) which is directed to the treatment of congestion and its management in Chinese production. As the study by Deng (2003) shows, congestion is used (or at least tolerated) by the Chinese government in order to deal with the problem of providing employment for a huge labor force – with some 16,000,000–18,000,000 new entrants every year. However, congestion is not confined to such examples. It also appears in other contexts such as when the inflow from one department leads to large excesses of "in process" inventories in another department of the same company. See Balakrishnan and Soderstrom (2000). In either case it may be possible to improve the management of this inventory flow and thereby improve output performances along lines like those indicated by the movement from E to F in Fig. 7.2.

We now note that we are dealing only with congesting inputs, partly because that is where the main managerial interest lies, and partly because it is not clear what is to be meant by congesting outputs. The topic of output congestion is dealt with mathematically by Färe et al. (1985, 1994). However, their treatment is restricted to a formal development without reference to any actual example of possible occurrences.

Models for distinguishing between different characterizations of inefficiency such as "congestion" and "technical inefficiency" will shortly be introduced. These concepts will also be developed further as the different applications we consider are brought into view. However, before undertaking to do this, we will first briefly review the DEA literature that has treated congestion in order to obtain additional perspective on the topics we will be covering.

7.2 Comparison of Two Literatures on Congestion

Congestion has been an under-researched topic in western economics partly because Stigler (1976), a Nobel Laureate economist, questioned whether "congestion" as a topic of research should have any place in economics in his review of the "X-Efficiency" concept of Leibenstein (1966, 1976). However, after a long period of neglect in the economics literature, Färe and Svensson (1980) initiated new research in this area by reformulating some of the concepts associated with congestion. Färe and Grosskopf (1983) then gave these concepts an operational form. Färe et al. (1985) subsequently finalized the models (and methods of analysis) that they used to analyze congestion and accorded them a form that would now be identified with DEA. This approach was the only one available in the DEA literature and was therefore employed in all of the research into congestion in the numerous applications that were then undertaken. Cooper, Thompson and Thrall (CTT 1996), however, formulated an alternative approach which has also begun to see various extensions and applications.

Interest in the development of alternative approaches has now begun to result in additions and extensions to CTT (1996) in a variety of ways. This has been advantageous because new alternatives provide perspective on shortcomings as well as advantages in the use of existing models. This was exhibited, for instance, in the exchanges between Färe and Grosskopf (1998), Färe and Grosskopf (2000a, b) and Brockett et al. (1998) and Cooper et al. (2000, 2001d). Shortcomings in the Färe, Grosskopf and Lovell approach were then identified by Cooper et al. (2000, 2001c) in a manner that led to the exchanges between Cherchye et al. (2001) and Cooper et al. (2001b).

Many new applications have been reported in various fields. Numerous references which involve such applications may be found in the bibliography for DEA compiled by Seiford (1994). This bibliography has now been extended and incorporated in a CD-ROM that accompanies the textbook by Cooper et al. (2007). Additional models have also come into being which we survey in the sections that follow after we start with FGL (Färe et al. 1985) because it has been the longest standing and most used approach to congestion in the DEA literature.

7.3 Färe, Grosskopf, and Lovell (FGL) Approach

The FGL approach proceeds in two stages. The first stage uses an "input-oriented" model as follows (Färe et al. 1985):

$$
\begin{aligned}
&\theta^* = \min \theta \\
&\text{subject to} \\
&\theta x_{io} \geq \sum_{j=1}^{n} x_{ij}\lambda_j, \quad i = 1, 2, \ldots, ml \\
&y_{ro} \leq \sum_{j=1}^{n} y_{rj}\lambda_j, \quad r = 1, 2, \ldots, s \\
&\lambda_j \geq 0, \qquad\qquad j = 1, 2, \ldots, n,
\end{aligned} \tag{7.1}
$$

where $j = 1, \ldots, n$ indexes the set of DMUs (decision making units) which are of interest. Here, x_{ij} is the observed amount of input $i = 1, \ldots, m$ used by DMU_j and y_{ro} is the observed amount of output $r = 1, \ldots, s$ produced by DMU_j. The x_{io} and y_{ro} represent the amounts of inputs $i = 1, \ldots, m$ and outputs $r = 1, \ldots, s$ associated with DMU_o, where DMU_o is the $\text{DMU}_j = \text{DMU}_o$ to be evaluated relative to all DMU_j (including itself) via (7.1). The objective is to minimize all of the inputs of DMU_o in the proportion θ^* where, because the $x_{io} = x_{ij}$ and $y_{ro} = y_{rj}$ for $\text{DMU}_j = \text{DMU}_o$ appear on both sides of the constraints in (7.1), the optimal $\theta = \theta^*$ does not exceed unity and the nonnegativity of the λ_o, x_{ij} and y_{ij} implies that the value of θ^* will not be negative under the optimization in (7.1). Hence,

$$0 \leq \text{Min } \theta = \theta^* \leq 1. \tag{7.2}$$

We now have the following definition of technical efficiency and inefficiency.

Definition 7.2. FGL Technical Efficiency:

1. Technical efficiency is achieved by DMU_o if and only if $\theta^* = 1$.
2. Technical inefficiency is present in the performance of DMU_o if and only if $0 \leq \theta^* < 1$.

This definition ignores the possible presence of nonzero slacks even when the solution of (7.1) shows them to be present. We therefore say that this definition refers to "weak" technical efficiency. This is the term used in the operation research literature. In the economics literature, it is referred to as the assumption of "strong disposal." In any case, FGL then go on to the following second stage model:

$$\begin{aligned} &\beta^* = \min\ \beta \\ &\text{subject to} \\ &\beta x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j, \quad i = 1, 2, \ldots, m \\ &y_{ro} \leq \sum_{j=1}^{n} y_{rj}\lambda_j, \quad r = 1, 2, \ldots, s \\ &\lambda_j \geq 0, \qquad\qquad j = 1, 2, \ldots, n. \end{aligned} \tag{7.3}$$

Note that the first $i = 1, \ldots, m$ inequalities in (7.1) are replaced by equations in (7.3). Thus slack is not possible in the inputs. The fact that only the output can yield nonzero slack is then referred to as "weak disposal" by Färe et al. (1985).

Now we note that (7.3) is more restricted than (7.1) by virtue of replacing inequalities with equations. Hence, we have $0 \leq \theta^* \leq \beta^*$. FGL use this property to develop a "measure" of congestion,

$$0 \leq C(\theta^*, \beta^*) = \frac{\theta^*}{\beta^*} \leq 1. \tag{7.4}$$

Table 7.1 Congestion example

DMU	Output y	Input x_2	Input x_1
1	2	1	2
2	2	2	2
3	2	2	1
4	2	1	3
5	2	1	4
6	2	3	1.25
7	2	4	1.25

Source: Färe et al. (1985, p. 160)

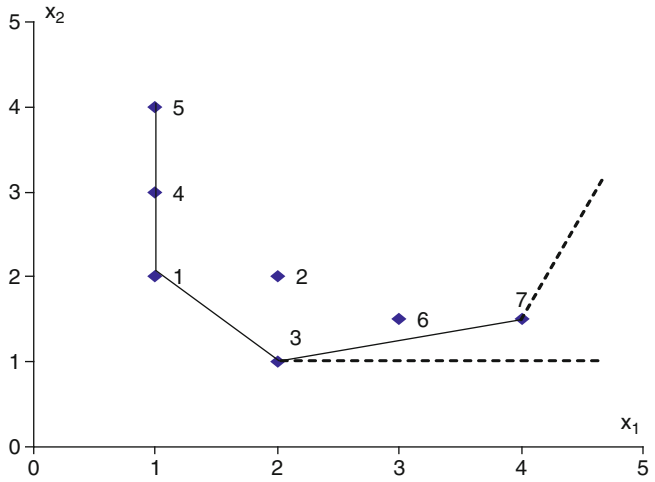


Fig. 7.3 Representation of points in Table 7.1

Combining models (7.1) and (7.3) in a two-stage manner, FGL utilize this measure to identify congestion in terms of the following conditions:

1. Congestion is identified as present in the performance of DMU$_o$ if and only if

$$C(\theta^*, \beta^*) < 1. \tag{7.5}$$

2. Congestion is identified as not present in the performance of DMU$_o$ if and only if

$$C(\theta^*, \beta^*) = 1.$$

Table 7.1, taken from Färe et al. (1985 p. 160), provides a numerical example which involves 7 DMUs in a two input and one output case.

To help explain this approach, we can also utilize Fig. 7.3, taken from FGL (1985) which uses the x_1, x_2 values in Table 7.1 to represent these DMUs geometrically in a Cartesian coordinate system. The line segments in Fig. 7.3 form what is called an "isoquant" (level line) in economics. These segments are obtained by running a plane through the three-dimensional production surface at the level $y = 2$

Table 7.2 Input efficiency and congestion in the example of Table 7.1

DMU	θ^*	β^*	$\frac{\theta^*}{\beta^*}$
1	1.0	1.0	1.0
2	0.75	0.75	1.0
3	1.0	1.0	1.0
4	1.0	1.0	1.0
5	1.0	1.0	1.0
6	0.8	0.86	0.93
7	0.8	1.0	0.8

Source: Färe et al. (1985, p. 76)

recorded in the output (=y) column of Table 7.1. After projecting these results down into a two-dimensional representation we obtain Fig. 7.3 in which all of the points shown, including those not on the isoquant represent x_1, x_2 coordinate values associated with the output values of $y = 2$ recorded in Table 7.1.

The line segment connecting points 3 and 7 is referred to as "backward bending" by FGL. To see what this means and how it is associated with congestion, consider point 7. Moving to the left along the level line from 7 to 3 is associated with reductions in both inputs while maintaining output at $y = 2$. This suggests the presence of congestion which is further confirmed by noting that horizontal movement to the left from 7 represents a reduction in x_1 that provides access to an output level on the surface of the PPS which is greater than $y = 2$. Hence, as required for congestion to be present, a reduction in input is thereby associated with an increase in the output that is maximally possible.

To see how FGL use models (7.1) and (7.3) and the measure supplied by (7.4) to determine whether congestion is present, we return to the data in Table 7.1 which provides the coordinates of the points in Fig. 7.1. Applying models (7.1) and (7.3) in the two-stage manner we have already described yields, the values shown in Table 7.2.

In accordance with (7.5), only DMUs 6 and 7, where the value for $C(\theta^*, \beta^*) = \theta^*/\beta^*$ is less than 1, display congestion in their performance. All of the other DMUs are identified as at least "weakly efficient" except DMU_2 which is technically inefficient, because $\theta^* < 1$, but nevertheless has a value of $\theta^*/\beta^* = 1$.

To relate the solutions in Table 7.2 to the diagram in Fig. 7.3 we focus on the points for DMUs 6 and 7, which are the only ones that satisfy the condition for congestion specified in (7.5). Using the coordinate for $x_2 = 1.25$, as given for DMUs 6 and 7 in the last column of Table 7.1 and coupling this value with $\theta^* = 0.8$ in Table 7.2, we obtain $0.8(1.25) = 1.00$. This positions both of these points on the horizontal line emanating from DMU_3 with coordinates $(x_1, x_2) = (2, 1)$ in Fig. 7.1. Evidently DMU_3 is the only point on this line which is efficient so, at this point, we must have nonzero slack. Therefore, there must be slack removal in input 1 for DMUs 6 and 7. To confirm this, we note that the intersection point with DMU_3 of this coordinate for DMU_6 is $0.8(3) = 2.4$ and for DMU_7 it is $0.8(4) = 3.2$. Thus, 0.4 unit of slack in input 1 must be removed for DMU_6 and 1.2 units must be removed for DMU_7 in order to obtain coincidence with DMU_3 at $(x_1, x_2) = (2, 1)$. Hence, the assumption of "weak efficiency" must be dropped in order to fully portray this solution.

Having examined the solution for (7.1) in terms of θ^* we next turn to the solution for β^* in (7.3). The equality condition for β^* in (7.3) requires solutions to lie on the line segment connecting the points for DMU_3 and DMU_7 and the algebraic expression which corresponds to this line is $x_2 = 0.75 + 0.125x_1$. As is evident from Fig. 7.3, DMU_7 is already on this line, which is confirmed by substituting the value $x_1 = 4$ in this algebraic expression to obtain $x_2 = 1.25$. For DMU_6 we use $\beta^* = 0.86$ to obtain the pertinent coordinates $\hat{x}_1 = 0.86(3) = 2.58$ and $\hat{x}_2 = 0.86(1.25) = 1.075$ for substitution in this algebraic expression to obtain $1.075 \approx 0.75 + 0.125(2.58)$ where we have used the approximation symbol "≈" to allow for roundoff error in the value of $\beta^* = 0.86$ reported by FGL in Table 7.2.

The rationale underlying the use of the measure in (7.4) in the manner specified by (7.5) to determine whether congestion is present can now be made clear from the following considerations. The horizontal broken line represents a boundary (=frontier) of the PPS as determined from (7.1). (There will always be such a boundary.) Hence, a value of $\theta^*/\beta^* < 1$ can occur only if the isoquant generated from some observation such as DMU_7 is "backward bending."

We now return to the issue of nonzero slacks that was noted in connection with our discussion of the θ^* value obtained from (7.1). The presence of such nonzero slack plays a critical role in the FGL development not only in this case but also in the case of movement to the right from this backward bending segment (such movements are associated with output reductions) or movements above this segment (such movements are associated with output increases). FGL therefore associate congestion with mix inefficiencies. For a definition of mix and its uses, see the discussion in Cooper et al. (1999). This condition is limitational rather than general, however, since one can readily supply examples where it does not hold. For instance, returning to the example of an underground mine one can have situations in which congestion is due to too many miners and too much equipment. Withdrawing these two inputs without changing their proportion can therefore serve to eliminate this congestion without altering the mix proportion.

Troubles are also encountered in cases where one wants to identify the sources and amounts of congesting inputs. The methods supplied for treating this problem, as described by FGL (1994, pp. 76–77), involve solving for each of the pertinent partition of the m inputs. This can be onerous, so when the identification of such inputs and their congesting amounts is of interest it is better to turn to one of the other models we shall shortly describe.

Other deficiencies are described in Cooper et al. (2001b, c), who show the following: (1) the FGL approach can show congestion to be present when this is not the case and (2) it can fail to show congestion to be present even when this is the case.

7.4 Cooper, Thompson, and Thrall (CTT) Approach

Cooper et al. (1996) introduced another model which was extended by Brockett et al. (1998) in their study of congestion in Chinese production. See the further developments on the use of these results for policy guidance in

Cooper et al. (2001a). This alternate approach also proceeds in a two-stage manner with the following "output-oriented" model used in the first stage.

$$\max \varphi_o + \varepsilon \left(\sum_{r=1}^{s} s_r^+ + \sum_{i=1}^{m} s_i^- \right)$$

subject to

$$x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-, \qquad i = 1, 2, \ldots, m$$

$$\varphi_o y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+, \quad r = 1, 2, \ldots, s$$

$$1 = \sum_{j=1}^{n} \lambda_j$$

$$0 \le \lambda_j, s_r^+, s_i^-, \qquad j = 1, \ldots, n; \quad i = 1, \ldots, m; \quad r = 1, \ldots, s. \tag{7.6}$$

Comparison with the first stage of the FGL approach in (7.3) shows the following three differences: (a) the convexity condition $\sum \lambda_j = 1$ in (7.6) is added to (7.3); (b) the objective in (7.6) is for an "output-oriented" model whereas an "input orientation" is used in the objective of (7.3); (c) the model (7.6) introduces slacks into the objective multiplied by a value $\varepsilon > 0$ which insures that some nonzero slack is not overlooked, possibly in an alternative optimum, as described for (1.2) in Chap. 1.

The second stage of this CTT approach, as taken from Brockett et al. (1998), can be represented as follows:

$$\begin{aligned}
&\text{Max} \sum_{i=1}^{m} \delta_i^- \\
&\text{Subject to} \\
&\hat{x}_{io} = \sum_{j=1}^{n} x_{ij}\hat{\lambda}_j - \delta_i^-, \quad i = 1, 2, \ldots, m \\
&\hat{y}_{ro} = \sum_{j=1}^{n} y_{rj}\hat{\lambda}_j, \qquad r = 1, 2, \ldots, s \\
&1 = \sum_{j=1}^{n} \hat{\lambda}_j \\
&s_i^{-*} \ge \delta_i^-, \quad i = 1, 2, \ldots, m \\
&0 \le \hat{\lambda}_j, \delta_i^-, \quad \forall i, j,
\end{aligned} \tag{7.7}$$

where $\hat{y}_{ro}$, $\hat{x}_{io}$ represent coordinates of a point on the efficiency frontier. These $\hat{y}_{ro}$, $\hat{x}_{io}$ are the coordinates of the point used to evaluate DMU_o. They are obtained from the solution of (7.6) and have the following values:

$$\begin{aligned}
\hat{x}_{io} &= x_{io} - s_i^{-*}, \qquad i = 1, 2, \ldots, m \\
\hat{y}_{ro} &= \varphi_o^* y_{ro} + s_r^{+*}, \quad r = 1, 2, \ldots, s.
\end{aligned} \tag{7.8}$$

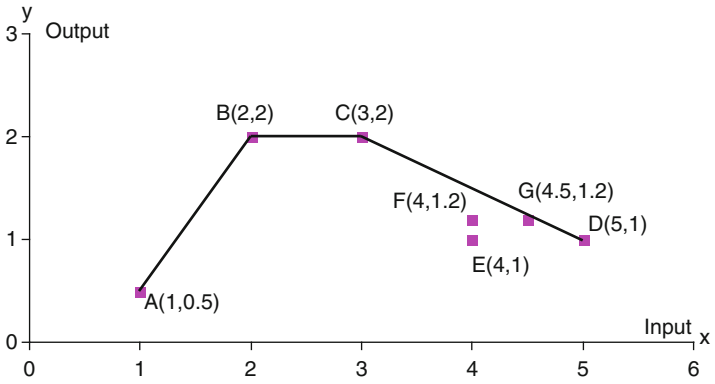


Fig. 7.4 A numerical example. Source: Brockett et al. (1998)

Remark. Comparison with (2.3) in Chap. 2 shows that (7.8) is also a projection formula which can be used in place of the input-oriented model projection associated with (2.3).

Notice that the inequalities implied for the inputs by the first $i = 1, \ldots, m$ constraints in (7.7) are reversed from the usual form as exhibited in (7.6) (cf. the change in sign for the slacks). The objective in (7.7) is to maximize the sum of the input slacks with the additional constraint $s_i^{-*} \geq \delta_i^-$ limiting each slack to the maximum value obtained in the preceding solution to (7.6). The difference may be represented as

$$s_i^{-c} = s_i^{-*} - \delta_i^{-*}, \quad i = 1, \ldots, m, \tag{7.9}$$

where for each $i = 1, \ldots, m$, δ_i^{-*} is obtained by solving (7.7) after s_i^{-*} has been subtracted from x_{io} as in (7.8). If desired, these slacks may be stated in ratio form relative to the observed x_{io}, as in Cooper et al. (2007), in order to obtain a measure of congestion which is invariant to the units of measure used. See (7.10) below.

These s_i^{-c} values, when positive, represent congesting amounts in each of the $i = 1, \ldots, m$ inputs while the $\delta_i^{-*} \geq 0$ represent corresponding technical inefficiency components. Thus, referring to s_i^{-*} as "total slack" (as obtained from (7.6)), we have, $s_i^{-*} = s_i^{-c} + \delta_i^{-*}$ $i = 1, \ldots, m$. That is, each "total slack," as obtained from (7.6) in stage one, is decomposed into two components via (7.7) in stage two. These two components consist of (1) ordinary technical inefficiency in amount δ_i^{-*} and (2) congesting amount s_i^{-c} as defined in (7.9).

7.4.1 A Numerical Example

Figure 7.4 can help to explain what is intended. Note, for instance, that point C with coordinates $(x, y) = (3, 2)$ is inefficient relative to B because B produced the same

2 units of output as C but did so with only 2 units of input. Hence, compared to B, C used an excess of 1 unit of input. However, there is no output reduction associated with this input excess, so the resulting inefficiency is "purely technical," i.e., no congestion is present at C.

While no evidence of congestion is present in the performance of C, all points to the right of C, do, in fact, exhibit congestion. For illustration, we apply models (7.6) and (7.7) to E in the two-stage manner that we just described. Application of (7.6) to the evaluation of E, with coordinate $(x, y) = (4, 1)$, yields the following model:

$$
\begin{aligned}
&\max \varphi + \varepsilon(s^+ + s^-)\\
&\text{subject to}\\
&1\varphi = 0.5\lambda_A + 2\lambda_B + 2\lambda_C + 1\lambda_D + 1\lambda_E + 1.2\lambda_F + 1.2\lambda_G - s^+\\
&4 = 1\lambda_A + 2\lambda_B + 3\lambda_C + 5\lambda_D + 4\lambda_E + 4\lambda_F + 4.5\lambda_G + s^-\\
&1 = \lambda_A + \lambda_B + \lambda_C + \lambda_D + \lambda_E + \lambda_F + \lambda_G\\
&0 \le \lambda_A, \ldots, \lambda_G, s^+, s^-.
\end{aligned}
$$

This has as its solution $\varphi^* = 2, \lambda_B^* = 1, s^{-*} = 2$. Thus, E is inefficient with both (a) $\varphi^* > 1$ and (b) some slacks are positive.

To ascertain whether congestion is present, we turn to (7.7) and use the results we have just secured to replace the preceding model with

$$
\begin{aligned}
&\max \delta^-\\
&\text{subject to}\\
&2 = 1\lambda_A + 2\lambda_B + 3\lambda_C + 5\lambda_D + 4\lambda_E + 4\lambda_F + 4.5\lambda_G - \delta^-\\
&2 = 0.5\lambda_A + 2\lambda_B + 2\lambda_C + 1\lambda_D + 1\lambda_E + 1.2\lambda_F + 1.2\lambda_G\\
&1 = \lambda_A + \lambda_B + \lambda_C + \lambda_D + \lambda_E + \lambda_F + \lambda_G\\
&2 \ge \delta^-\\
&0 \le \lambda_A, \ldots, \lambda_G, \delta^-.
\end{aligned}
$$

Here, $\hat{x} = 2$ and $\hat{y} = 2$, as represented on the left in this model, are obtained by using the preceding solution to obtain the values $\hat{x} = 2$, $\hat{y} = 2$ by means of (7.8). Similarly, $s^{-*} = 2$ (from the preceding solution) bounds the admissible values of δ^- via the last constraint.

The solution to the just stated problem is $\lambda_C^* = 1, \delta^{-*} = 1$, and all other variables at zero. Substitution of δ^{-*} in (7.9), therefore, gives $s^{-c} = 2 - 1 = 1$ as the congesting amount of this input that is identified in the performance of E. Hence, the total slack, at $s^{-*} = 2$, as obtained from the first stage, is decomposed in this solution to a value of $\delta^{-*} = 1$ for "purely technical inefficiency" and $s^{-c} = 1$ for the "congesting" part of this "total slack," so that, as required, $s_i^{-*} = s_i^{-c} + \delta_i^{-*}$.

All of this information, which is automatically available, can be related to Fig. 7.4 by noting that $s^{-*} = 2$ is the reduction in the $x = 4$ units of input used by E which is necessary to obtain coincidence with the input coordinate of the efficient point represented by B. At the same time, $s^{-c} = 1$ is the reduction in the input of E needed for coincidence with the input coordinate of C. Further, $\delta^{-*} = 1$ is the amount (here = one unit) of technical inefficiency that needs to be removed from the input of C to attain coincidence with the input of B. Finally, comparing the two units of output at C with the one unit of observed output for E shows the amount of reduction (=one unit) in output. As will be seen in Sect. 7.6, there is also a managerial inefficiency component to be accounted for in this output reduction, which is represented by the fact that output fails to achieve the boundary (=maximal output value) that is attainable with the 4 units of input used at E.

7.5 A Unified Additive Model

The FGL approach, as described earlier, uses a radial measure of inefficiency in both stages one and two. The CCT approach, however, uses a radial measure only in stage one. In stage two, on the other hand, CCT uses a modified version of an "additive model" which measures the amount of inefficiency in terms of "ℓ_1 measure." The "ℓ_1 measure" is also called the "city block measure" of distance in the operations research literature. A comprehensive treatment of ℓ_1 and other measures used in mathematics may be found in Appendix A of Charnes and Cooper (1961).

The publication, by Cooper, Seiford, and Zhu (CSZ 2007) replaces the mixture of (radial and ℓ_1) measures used in the CTT approach. The version of the additive model used by CSZ unifies matters by applying the same "ℓ_1 measure" in both stages. Hence, we now turn to this CSZ approach to describe how this is accomplished.

The first stage model in this approach is formulated by CSZ in the following manner:

$$\text{Max} \frac{\sum_{r=1}^{s} \frac{s_r^+}{y_{ro}}}{s} + \varepsilon \frac{\sum_{i=1}^{m} \frac{s_i^-}{x_{io}}}{m}$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io}, \quad i = 1, \ldots, m$$

$$\sum_{j=1}^{n} \lambda_j y_{ij} - s_r^+ = y_{ro}, \quad r = 1, \ldots, s$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j, s_i^-, s_r^+ \geq 0, \tag{7.10}$$

where $\varepsilon>0$ is the non-Archimedean element, smaller than any positive real number, that was described earlier in the discussion of (7.6). Also as discussed in association with (1.2) in Chap. 1 of this handbook, this use of $\varepsilon>0$ accords preemptive priority to maximizing $\sum_{r=1}^{s} s_r^+/y_{ro}$.

The results from (7.10) are used to form the following model:

$$\begin{aligned}
&\text{Max} \frac{\sum_{i=1}^{m} \frac{s_i^-}{x_{io}}}{m} \\
&\text{subject to} \\
&\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io}, \quad i = 1, \ldots, m \\
&\sum_{j=1}^{n} \lambda_j y_{ij} = \hat{y}_{ro}, \quad r = 1, \ldots, s \\
&\sum_{j=1}^{n} \lambda_j = 1 \\
&\lambda_j, s_i^- \geq 0,
\end{aligned} \tag{7.11}$$

where $\hat{y}_{ro} = y_{ro} + s_r^{+*}$ so that $\hat{y}_{ro}$ with the slacks, s_r^{+*}, representing output slack, $r = 1, \ldots, s$, is obtained from (7.10).

The solution of (7.11) yields a new set of maximal input slacks consistent with the thus adjusted outputs. We next "back out" the maximal inputs. This backing out is accomplished by means of the following modification of (7.7):

$$\begin{aligned}
&\text{Max} \frac{\sum_{i=1}^{m} \frac{\delta_i^-}{x_{io}}}{m} \\
&\text{subject to} \\
&\sum_{j=1}^{n} \lambda_j x_{ij} - \delta_i^- = \hat{x}_{io}, \quad i = 1, \ldots, m \\
&\sum_{j=1}^{n} \lambda_j y_{ij} = \hat{y}_{ro}, \quad r = 1, \ldots, s \\
&\sum_{j=1}^{n} \lambda_j = 1 \\
&\delta_i^- \leq s_i^{-*} \\
&\lambda_j, \delta_i^- \geq 0,
\end{aligned} \tag{7.12}$$

where $\hat{x}_{io} = x_{io} - s_i^{-*}$ with s_i^{-*} representing the optimal slacks obtained from (7.11) in the second stage optimization associated with $\varepsilon > 0$ in (7.10).

This returns us to (7.9) to obtain the values of $s_i^{-c} = s_i^{-*} - \delta_i^{-*} \geq 0$ where s_i^{-c} represents the amount of congestion in each input $i = 1, \ldots, m$. In many cases, the output reductions resulting from congestion will be apparent from inefficiency. For a formal development that will handle all cases, however, we now replace (7.11) with

$$
\begin{aligned}
&\text{Max} \frac{\sum_{r=1}^{s} \frac{\delta_i^+}{y_{ro}}}{m} \\
&\text{subject to} \\
&\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io}, \quad i = 1, \ldots, m \\
&\sum_{j=1}^{n} \lambda_j y_{ij} + \delta_r^+ = \hat{y}_{ro}, \quad r = 1, \ldots, s \\
&\sum_{j=1}^{n} \lambda_j = 1 \\
&\lambda_j, \delta_r^+ \geq 0,
\end{aligned}
\tag{7.13}
$$

where $\hat{y}_{ro}$ is defined as in (7.11) and the x_{io} are original data as in the inputs for (7.10) and (7.11). When the optimal solution for (7.10) is unique, the solution to (7.13) will simply reproduce the original data via $\hat{y}_{ro} - \delta_r^{+*} = y_{ro}$. When not unique, however, other possibilities may be present.

7.6 Estimating the Output Effects of Congestion

Figure 7.5 is a modified version of Fig. 7.4 that can help us to determine the output effects of congestion. As shown in the case of Fig. 7.4 we can use (7.7) and (7.9) to determine the congesting amount of input as $s^{-c} = 1$, which represents the difference between $x_{\mathrm{E}} = 4$ at E and $x_{\mathrm{H}} = 3$ at H, the point where technical inefficiency gives way to congestion. Thus, the amount of congesting input is $x_{\mathrm{E}} - x_{\mathrm{H}} = 1$. However, the decrease in output is given by the coordinate value of $y_{\mathrm{E}'} = 1.5$ at E', according to Definition 7.1, and not by the coordinate value of $y_{\mathrm{E}} = 1$, since the latter reflects a managerial inefficiency as well as a congestion component.

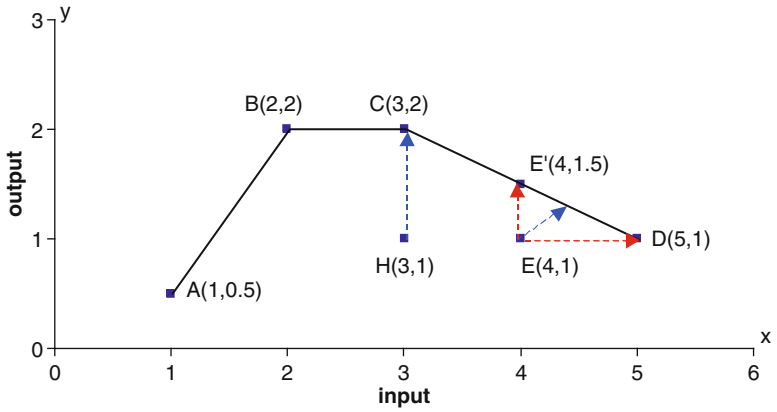


Fig. 7.5 Congestion and its effects

To estimate the amount of managerial inefficiency we use the following version of our output-oriented "additive" model as taken from Cooper et al. (2001a):

$$\begin{aligned}
&\max \sum_{r=1}^{s} s_r^+ \\
&\text{subject to} \\
&y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+, \quad r = 1, 2, \ldots, s; \\
&x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j - s_i^-, \quad i = 1, 2, \ldots, m; \\
&1 = \sum_{j=1}^{n} \lambda_j \\
&0 \le \lambda_j, s_i^-, s_r^+, \quad \forall i, j, r.
\end{aligned} \tag{7.14}$$

To see what is involved, we note that the input (like the output) constraints take the form $x_{io} < \sum x_{ij}\lambda_j$, with the slacks $s_i^- \ge 0$ used to replace the inequalities with equivalent equations. Hence, the solution values of the inputs, just like the outputs, all need to equal or exceed these observed values, x_{io}, for this DMU$_o$, so that no input reductions are allowed in this model. In this adaptation of the additive models, the outputs are to be maximized without reducing any inputs. Formally, in (7.14) the inputs may be increased provided this does not reduce the maximum output values that might otherwise be available.

To illustrate what is involved we use the coordinate values in Fig. 7.5 and employ the model in (7.14) to evaluate E as DMU$_o$. This produces

$$
\begin{aligned}
&\max\ s^{+}\\
&\text{subject to}\\
&1 = 0.5\lambda_{\mathrm{A}} + 2\lambda_{\mathrm{B}} + 2\lambda_{\mathrm{C}} + 1\lambda_{\mathrm{D}} + 1\lambda_{\mathrm{E}} + 1\lambda_{\mathrm{H}} - s^{+}\\
&4 = 1\lambda_{\mathrm{A}} + 2\lambda_{\mathrm{B}} + 3\lambda_{\mathrm{C}} + 5\lambda_{\mathrm{D}} + 4\lambda_{\mathrm{E}} + 3\lambda_{\mathrm{H}} - s^{-}\\
&1 = \lambda_{\mathrm{A}} + \lambda_{\mathrm{B}} + \lambda_{\mathrm{C}} + \lambda_{\mathrm{D}} + \lambda_{\mathrm{E}} + \lambda_{\mathrm{H}}\\
&0 \le \lambda_{\mathrm{A}}, \ldots, \lambda_{\mathrm{H}}, s^{+}, s^{-}.
\end{aligned}
$$

This problem has as its solution

$$\lambda_{\mathrm{C}}^{*} = \lambda_{\mathrm{D}}^{*} = 1/2, \quad s^{+*} = 1/2,$$

and all other variables zero. Hence, we have $y_o + s^{+*} = 1 + 1/2 = 1.5$, so the coordinates of $E = (4, 1)$ give way to $E' = (4, 1.5)$. Thus, without any reduction in this congesting amount of input we get an increase in output. Finally, to show that this point is on the line segment connecting C and D in Fig. 7.5, we need merely substitute $x_{\mathrm{E}'} = 4$ in $y = 7/2 - 1/2 \times x$, the equation that corresponds to this line segment, to obtain $y = 7/2 - 4/2 = 3/2$ and thereby confirm the previously obtained 1/2 unit increase that replaces $y = 1$ with $y = 3/2$ via the just displayed solution.

We can now use the following formulas to estimate the output reductions due to congestion

$$\hat{y}_{ro} - \breve{y}_{ro} \ge 0, \quad r = 1, ..., s, \tag{7.15}$$

where the $\hat{y}_{ro}$ are obtained from (7.8) or (7.14) – depending on which model was used – and the $\hat{y}_{ro}$ are obtained from the solution to (7.14) via the following formulas

$$\breve{y}_{ro} = y_{ro} + s_r^{+*}, \quad r = 1, ..., s, \tag{7.16}$$

with the s_r^{+*} obtained from (7.14) after modifying its constraints by replacing $x_{io} \le \sum x_{ij}\lambda_j$ with $x_{io} = \sum x_{ij}\lambda_j, \quad i = 1, 2, \ldots, m.$

For illustration we note that $s^{+*} = 1/2$ is the only nonzero slack in the solution to the preceding numerical example and that $y = 1$. Hence, using (7.16), $\breve{y}_{ro} = y + s^{+*} = 1\frac{1}{2}$ is found to be the ordinate for E' in Fig. 7.5. We next subtract this result from the value of $\hat{y} = 2$ which was obtained by applying (7.7) and (7.8) to the evaluation of E. This gives $\hat{y} - \breve{y} = 2 - 1\frac{1}{2} = 1/2$ as the output lost because of the congesting input.

Finally, to see that the inequality represented in (7.15) is justified we note that the optimization obtained from the modification of (7.14) is restricted to a very small subset of the region available to the models used to obtain the $\hat{y}_{ro}$. Reference to the definition of congestion given in (7.1) then shows that no increases in any

output can accompany the decreases in some outputs as a result of the congestion associated with some of the inputs. This condition is necessary to distinguish the effects of congestion from the case where output substitutions with accompanying changes in inputs are made in which some output decreases are needed to secure increases in other outputs. Thus, if $\hat{y} - \breve{y} < 0$ for any $r = 1, \ldots, s$, then "substitution" rather than "congestion" is present.

7.7 Extensions

We have now covered the standard approaches used for treating congestion in DEA. Progress continues to be made with other models that have recently appeared in the literature. We do not cover these models in detail because experience with their use has not accumulated to a point where this appears to be warranted. We simply refer to them with a few brief remarks and provide references for readers who wish to study them further.

One such model due to Cooper et al. (2002a) is referred to as the "one-model approach" because it replaces the use of two models, as in the other approaches we have covered. This model can be written

$$\begin{aligned}
&\max \varphi + \varepsilon \left(\sum_{r=1}^{s} s_r^{+} - \sum_{i=1}^{m} s_i^{-c} \right) \\
&\text{subject to} \\
&\varphi y_{r0} - \sum_{j=1}^{n} y_{rj} \lambda_j + s_r^{+} = 0, \quad r = 1, \ldots, s, \\
&\sum_{j=1}^{n} x_{ij} \lambda_j + s_i^{-c} = x_{io}, \quad i = 1, \ldots, m, \\
&\sum_{j=1}^{n} \lambda_j = 1, \\
&0 \leq \lambda_j; s_i^{-c}, \quad j = 1, \ldots, n; \quad i = 1, \ldots, m.
\end{aligned} \tag{7.17}$$

As can be seen, the constraints in this model are the same as for (7.6), but (a) the objective of (7.6) is modified by replacing $+ s_i^{-}$ with $- s_i^{-c}$ for each $i = 1, \ldots, m$ and (b) the input constraints replace the usual slacks s_i^{-} with s_i^{-c}.

A series of theorems proved by Cooper et al. (2002a) are followed by examples that guide the use of (7.17). We do not reproduce these theorems here, however, other than to note that an optimum with $s^{-c*} > 0$ is required to show that congestion is present in the performance of the DMU_o being evaluated. However, access to the decomposition represented in (7.9) is lost so that identification of technical inefficiency is not available. Other information on efficient vs. inefficient behavior

is also absent from this "one-model" approach. Hence, if this information is wanted, recourse may be made to one of the models described earlier in this chapter.

This is, of course, not the end of the line for further openings to research in congestion and its management. Brockett et al. (2004) extend the study of congestion to include other ways to improve its management. This includes providing ways to estimate tradeoffs between inputs and outputs that extend the use of concepts such as "marginal rates of transformation" (from economics) in ways that can further improve the output values. This includes, for instance, the use of such estimates to assign congesting inputs to different plants or to different departments in order to further reduce its effects on outputs. This could prove useful to the Chinese government, for instance, by allowing it to deal with its huge and rapidly growing labor force in ways that can reduce the diminutions in output that might otherwise occur when its labor assignments result in congestion.

Finally, we note that work on introducing stochastic elements into the treatment of congestion, and other parts of DEA, has also begun to appear. See Cooper et al. (2002b, 2004) who treat both technical inefficiencies and congestion with chance-constrained programming models. See also Chap. 9 in this handbook for other uses of chance constrained-programming in DEA.

References

Balakrishnan R, Soderstrom NS. The cost of system congestion: Evidence from the health care sector. J Manage Account Res. 2000;12:97–114.

Brockett PL, Cooper WW, Deng H, Golden LL, Ruefli TW. Using DEA to identify and manage congestion. J Prod Anal. 2004;22(2):207–26.

Brockett PL, Cooper WW, Shin HC, Wang Y. Inefficiency and congestion in Chinese production before and after the 1978 economic reforms. Socioecon Plann Sci. 1998;32:1–20.

Charnes A, Cooper WW. Management models and industrial applications of linear programming. New York: John Wiley and Sons, Inc; 1961.

Cherchye L, Kuosmanen T, Post T. Alternative treatments of congestion in DEA: a rejoinder to Cooper, Gu and Li. Eur J Oper Res. 2001;132:75–80.

Cooper WW, Deng H, Gu B, Li S, Thrall RM. Using DEA to improve the management of congestion in Chinese industry (1981–1997). Socioecon Plann Sci. 2001a;35:1–16.

Cooper WW, Deng H, Huang Z, Li SX. A one-model approach to the analysis of congestion in Data Envelopment Analysis. Socioecon Plann Sci. 2002a;36:231–8.

Cooper WW, Deng H, Huang Z, Li SX. Chance Constrained Programming approaches to technical efficiencies and inefficiencies in stochastic data envelopment analysis. J Oper Res Soc. 2002b;53:1–10.

Cooper WW, Gu B, Li S. Alternative treatments of congestion in DEA: a response to the Cherchye, Kuosmanen and Post critique. Eur J Oper Res. 2001b;132:81–7.

Cooper WW, Gu B, Li S. Comparisons and evaluations of alternative approaches to evaluating congestion in DEA. Eur J Oper Res. 2001c;32(1):1–13.

Cooper WW, Deng H, Huang Z, Li SX. Chance Constrained Programming approaches to congestion in stochastic data envelopment analysis. Eur J Oper Res. 2004;155(2):487–501.

Cooper WW, Park KS, Pastor JT. RAM: a range adjusted measure of inefficiency for use with additive models and relations to other models and measures in DEA. J Prod Anal. 1999;11:5–42.

Cooper WW, Seiford LM, Tone K. A comprehensive text with uses: data envelopment analysis, example applications, references and DEA-Solver Software. 2nd ed. Norwell: Kluwer Academic Publishers; 2007.
Cooper WW, Seiford LM, Zhu J. A unified additive model approach for evaluating inefficiency and congestion with associated measures in DEA. Socioecon Plann Sci. 2000;34:1–25.
Cooper WW, Seiford LM, Zhu J. Slacks and congestion: a response to comments by Färe and Grosskopf, Socioecon Plann Sci. 2001d;35:1–11.
Cooper WW, Thompson RG, Thrall RM. Introduction: extensions and new developments in DEA. Ann Oper Res. 1996;66:3–45.
Deng HH. Congestion and its management in Chinese production. Ph.D. Thesis. The Red McCombs School of Business, University of Texas at Austin: Austin, Texas. Also available from University Microfilms, Inc.: Ann Arbor, MI; 2003.
Färe R, Grosskopf S. Congestion: a note. Socioecon Plann Sci. 1998;33:21–3.
Färe R, Grosskopf S. Congestion: a response. Socioecon Plann Sci. 2000a;34:35–50.
Färe R, Grosskopf S. Measuring congestion in production. Zeitschrift Für Nationalökonomie. 1983;43:251–71.
Färe R, Grosskopf S. Slacks and congestion: a comment. Socioecon Plann Sci. 2000b;34:27–33.
Färe R, Grosskopf S, Lovell CAK. Production frontiers. Cambridge: Cambridge University Press; 1994.
Färe R, Grosskopf S, Lovell CAK. The measurement of efficiencies of production. Boston: Kluwer-Nihoff Publishing; 1985.
Färe RS, Svensson L. Congestion of factors of production. Econometrica. 1980;48:1743–53.
Leibenstein H. Allocative efficiency vs. X-Efficiency. Am Econ Rev. 1966;56:392–415.
Leibenstein H. Beyond economic man. Cambridge: Harvard University Press; 1976.
Samuelson PA. Foundations of economics. Cambridge: Harvard University Press; 1947.
Seiford LM. References. In: Charnes A, Cooper WW, Lewin AY, editors. Data envelopment analysis: theory, methodology and applications. Kluwer Academic Publisher: Norwell; 1994.
Stigler GJ. The X-istence of X-efficiency. Am Econ Rev. 1976;66:213–6.
Varian H. Microeconomic analysis. 2nd ed. New York: W.W. Norton, Inc.; 1984.

Chapter 8
Slacks-Based Measure of Efficiency

Kaoru Tone

Abstract There are two types of models in DEA: radial and nonradial. Radial models are represented by the CCR (Charnes–Cooper–Rhodes) model. Basically, they deal with proportional changes of inputs or outputs. On the other hand, nonradial models, e.g., the slacks-based measure of efficiency (SBM) model, handle input or output slacks directly, and do not assume proportional changes of inputs or outputs. In this chapter, we introduce the SBM model and its extensions.

Keywords Slacks-based measure • Weighted-SBM • Super-SBM • Epsilon-based measure

8.1 Introduction

There are two types of models in DEA: radial and nonradial. Radial models are represented by the CCR (Charnes–Cooper–Rhodes) model. Basically, they deal with proportional changes of inputs or outputs. As such, the CCR score reflects the proportional maximum input (output) reduction (expansion) rate which is common to all inputs (outputs). However, in real-world businesses, not all inputs (outputs) behave in the proportional way. For example, if we employ labor, materials, and capital as inputs, some of them are substitutional and do not change proportionally. Another shortcoming of the radial models is the neglect of slacks in reporting the efficiency score. In many cases, we find a lot of remaining nonradial slacks. So, if these slacks have an important role in evaluating managerial efficiency, the radial approaches may mislead the decision when we utilize the efficiency score as the only index for evaluating performance of DMUs.

K. Tone (✉)
National Graduate Institute for Policy Studies, 7-22-1 Roppongi, Minato-ku,
Tokyo 106-8677, Japan
e-mail: tone@grips.ac.jp

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_8,

In contrast, the nonradial slacks-based measure of efficiency (SBM) models put aside the assumption of proportionate changes in inputs and outputs, and deal with slacks directly. This may discard varying proportions of original inputs and outputs. The SBM models are designed to meet the following two conditions.

1. Units invariant: The measure should be invariant with respect to the units of data.
2. Monotone: The measure should be monotone decreasing in each slack in input and output.

The rest of this chapter organized as follows. Section 8.2 introduces the SBM models in input-, output-, and nonoriented cases under the constant returns-to-scale assumption. We further observe the dual side of the model. We extend them to variable returns-to-scale (VRS) environment, weighted-SBM, and super-SBM models in Sect. 8.3. In Sect. 8.4, we deal with nonpositive data and several variants of the SBM models. A compromised model of the radial and nonradial model called "EBM" is introduced, too. Section 8.5 concludes this chapter.

8.2 The SBM Model

The SBM model was introduced by Tone (2001) (see also Pastor et al. 1999). It has three variations, i.e., input-, output-, and nonoriented. The nonoriented model indicates both input- and output-oriented.

Let the set of DMUs be $J = \{1, 2, \ldots, n\}$, each DMU having m inputs and s outputs. We denote the vectors of inputs and outputs for DMU$_j$ by $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{mj})^{\mathrm{T}}$ and $\mathbf{y}_j = (y_{1j}, y_{2j}, \ldots, y_{sj})^{\mathrm{T}}$, respectively. We define input and output matrices $\mathbf{X}$ and $\mathbf{Y}$ by

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \in R^{m \times n} \quad \text{and} \quad \mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) \in R^{s \times n}. \tag{8.1}$$

We assume that all data are positive, i.e., $\mathbf{X} > \mathbf{0}$ and $\mathbf{Y} > \mathbf{0}$.

8.2.1 Production Possibility Set

The production possibility set is defined using the nonnegative combination of the DMUs in the set J as:

$$P = \left\{ (\mathbf{x}, \mathbf{y}) \middle| \mathbf{x} \geq \sum_{j=1}^{n} \lambda_j \mathbf{x}_j,\ \mathbf{0} \leq \mathbf{y} \leq \sum_{j=1}^{n} \lambda_j \mathbf{y}_j,\ \boldsymbol{\lambda} \geq \mathbf{0} \right\}. \tag{8.2}$$

$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)^{\mathrm{T}}$ is called the intensity vector.

The inequalities in (8.2) can be transformed into equalities by introducing slacks as follows:

$$\begin{aligned} \mathbf{x} &= \sum_{j=1}^{n} \lambda_j \mathbf{x}_j + \mathbf{s}^-, \\ \mathbf{y} &= \sum_{j=1}^{n} \lambda_j \mathbf{y}_j - \mathbf{s}^+, \\ \mathbf{s}^- &\geq \mathbf{0}, \quad \mathbf{s}^+ \geq \mathbf{0}, \end{aligned} \tag{8.3}$$

where $\mathbf{s}^- = (s_1^-, s_2^-, \ldots, s_m^-)^{\mathrm{T}} \in R^m$ and $\mathbf{s}^+ = (s_1^+, s_2^+, \ldots, s_s^+)^{\mathrm{T}} \in R^s$ are, respectively, called input and output slacks.

8.2.2 Input-Oriented SBM

In order to evaluate the relative efficiency of $\mathrm{DMU}_o = (\mathbf{x}_o, \mathbf{y}_o)$, we solve the following linear program. This process is repeated n times for $o = (1, \ldots, n)$.

[SBM-I-C] (Input-oriented SBM under constant returns-to-scale assumption)

$$\begin{aligned} &\rho_I^* = \min_{\lambda, \mathbf{s}^-, \mathbf{s}^+} 1 - \frac{1}{m} \sum_{i=1}^{m} \frac{s_i^-}{x_{io}}, \\ &\text{subject to} \\ &x_{io} = \sum_{j=1}^{n} x_{ij} \lambda_j + s_i^- \quad (i = 1, \ldots, m), \\ &y_{ro} = \sum_{j=1}^{n} y_{rj} \lambda_j - s_r^+ \quad (r = 1, \ldots, s), \\ &\lambda_j \geq 0 \ (\forall j), \quad s_i^- \geq 0 \ (\forall i), \quad s_r^+ \geq 0 \ (\forall r). \end{aligned} \tag{8.4}$$

ρ_I^* is called SBM-input-efficiency.

Proposition 8.1. ρ_I^* is units invariant, i.e., it is independent of the units in which the inputs and outputs are measured.

Let an optimal solution of [SBM-I-C] be $(\boldsymbol{\lambda}^*, \mathbf{s}^{-*}, \mathbf{s}^{+*})$.

Definition 8.1 (SBM-Input-Efficient). A $\mathrm{DMU}_o = (\mathbf{x}_o, \mathbf{y}_o)$ is called SBM-input-efficient if $\rho_I^* = 1$ holds.

This means $\mathbf{s}^{-*} = \mathbf{0}$, i.e., all input slacks are zero. However, output slacks may be nonzero.

Definition 8.2 (Projection). Using an optimal solution $(\boldsymbol{\lambda}^*, \mathbf{s}^{-*}, \mathbf{s}^{+*})$, we define a projection of $\mathrm{DUM}_o = (\mathbf{x}_o, \mathbf{y}_o)$ by

$$(\overline{\mathbf{x}}_o, \overline{\mathbf{y}}_o) = (\mathbf{x}_o - \mathbf{s}^{-*}, \mathbf{y}_o + \mathbf{s}^{+*}). \tag{8.5}$$

Proposition 8.2. The projected DMU is SBM-input-efficient.

Definition 8.3 (Reference Set). We define a reference set R of $\mathrm{DUM}_o = (\mathbf{x}_o, \mathbf{y}_o)$ by

$$R = \left\{ j \middle| \lambda_j^* > 0,\ j \in J \right\}. \tag{8.6}$$

Thus, $(\mathbf{x}_o, \mathbf{y}_o)$ can be expressed as follows:

$$\begin{aligned} x_{io} &= \sum_{j \in R} x_{ij}\lambda_j^* + s_i^{-*} \quad (i = 1, \ldots, m), \\ y_{ro} &= \sum_{j \in R} y_{rj}\lambda_j^* - s_r^{+*} \quad (r = 1, \ldots, s). \end{aligned} \tag{8.7}$$

Proposition 8.3. DMUs in the reference set R to $(\mathbf{x}_o, \mathbf{y}_o)$ are SBM-input-efficient.

Proposition 8.4. The SBM-input-efficiency score is not greater than the CCR efficiency score. (See Tone 2001 for a proof.)

8.2.3 Output-Oriented SBM

The output-oriented SBM efficiency ρ_O^* of $\mathrm{DMU}_o = (\mathbf{x}_o, \mathbf{y}_o)$ is defined by

[SBM-O-C]

$$\begin{aligned} &\frac{1}{\rho_O^*} = \max_{\lambda, \mathbf{s}^-, \mathbf{s}^+} \ 1 + \frac{1}{s} \sum_{r=1}^{s} \frac{s_r^+}{y_{ro}}, \\ &\text{subject to} \\ &x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- \quad (i = 1, \ldots, m), \\ &y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ \quad (r = 1, \ldots, s), \\ &\lambda_j \geq 0\ (\forall j), \quad s_i^- \geq 0\ (\forall i), \quad s_r^+ \geq 0\ (\forall r). \end{aligned} \tag{8.8}$$

Let an optimal solution of [SBM-O-C] be $(\boldsymbol{\lambda}^*, \mathbf{s}^{-*}, \mathbf{s}^{+*})$.

Definition 8.4 (SBM-Output-Efficient). A $\mathrm{DMU}_o = (\mathbf{x}_o, \mathbf{y}_o)$ is called SBM-output-efficient if $\rho_O^* = 1$ holds.

This means $\mathbf{s}^{+*} = \mathbf{0}$, i.e., all output slacks are zero. However, input slacks may be nonzero.

Definition 8.5 (Projection). Using an optimal solution $(\boldsymbol{\lambda}^*, \mathbf{s}^{-*}, \mathbf{s}^{+*})$, we define a projection of $\mathrm{DUM}_o = (\mathbf{x}_o, \mathbf{y}_o)$ by

$$\left(\overline{\mathbf{x}}_o, \overline{\mathbf{y}}_o\right) = (\mathbf{x}_o - \mathbf{s}^{-*}, \mathbf{y}_o + \mathbf{s}^{+*}). \tag{8.9}$$

Proposition 8.5. The projected DMU is SBM-output-efficient.

8.2.4 Nonoriented SBM

Nonoriented or both-oriented SBM efficiency ρ_{IO}^* is defined by

[SBM-C]

$$\begin{aligned}
&\rho_{IO}^* = \min_{\boldsymbol{\lambda}, \mathbf{s}^-, \mathbf{s}^+} \frac{1 - (1/m)\sum_{i=1}^{m} (s_i^- / x_{io})}{1 + (1/s)\sum_{r=1}^{s} (s_r^+ / y_{ro})},\\
&\text{subject to}\\
&x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- \quad (i = 1, \ldots, m),\\
&y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ \quad (r = 1, \ldots, s),\\
&\lambda_j \geq 0\ (\forall j), \quad s_i^- \geq 0\ (\forall i), \quad s_r^+ \geq 0\ (\forall r).
\end{aligned} \tag{8.10}$$

Definition 8.6 (SBM-Efficient). A $\mathrm{DMU}_o = (\mathbf{x}_o, \mathbf{y}_o)$ is called SBM-efficient if $\rho_{IO}^* = 1$ holds.

This means $\mathbf{s}^- = \mathbf{0}$ and $\mathbf{s}^{+*} = \mathbf{0}$, i.e., all input and output slacks are zero.

[SBM-C] can be transformed into a linear program using the Charnes–Cooper transformation as follows:

[SBM-C-LP]

$$\begin{aligned}
&\tau^* = \min_{t, \boldsymbol{\Lambda}, \mathbf{S}^-, \mathbf{S}^+} t - \frac{1}{m}\sum_{i=1}^{m} \frac{S_i^-}{x_{io}},\\
&\text{subject to}\\
&1 = t + \frac{1}{s}\sum_{r=1}^{s} \frac{S_r^+}{y_{ro}},\\
&tx_{io} = \sum_{j=1}^{n} x_{ij}\Lambda_j + S_i^- \quad (i = 1, \ldots, m),\\
&ty_{ro} = \sum_{j=1}^{n} y_{rj}\Lambda_j - S_r^+ \quad (r = 1, \ldots, s),\\
&\Lambda_j \geq 0\ (\forall j), \quad S_i^- \geq 0\ (\forall i), \quad S_r^+ \geq 0\ (\forall r), \quad t > 0.
\end{aligned} \tag{8.11}$$

Table 8.1 Data of Example 1

DMU	(I) x_1	(I) x_2	(O) y_1	(O) y_2
A	4	3	1	2
B	14	6	2	6
C	24	3	3	12
D	20	2	2	6
E	48	4	4	16
F	50	7.5	5	30

Table 8.2 Score and rank of efficiency for Example 1

	CCR-I		SBM-I-C		SBM-O-C		SBM-C	
DMU	Score	Rank	Score	Rank	Score	Rank	Score	Rank
A	1	1	1	1	1	1	1	1
B	0.8085106	6	0.75	6	0.8067227	6	0.6923077	6
C	1	1	1	1	1	1	1	1
D	1	1	0.9	4	0.8571429	5	0.7714286	5
E	1	1	0.8333333	5	1	1	0.8333333	4
F	1	1	1	1	1	1	1	1

Let an optimal solution be $(\boldsymbol{\tau}^*, t^*, \boldsymbol{\Lambda}^*, \mathbf{S}^{-*}, \mathbf{S}^{+*})$. Then, we have an optimal solution of [SBM-C] as defined by

$$\rho^* = \tau^*, \quad \boldsymbol{\lambda}^* = \frac{\boldsymbol{\Lambda}^*}{t^*}, \quad \mathbf{s}^{-*} = \frac{\mathbf{S}^{-*}}{t^*}, \quad \mathbf{s}^{+*} = \frac{\mathbf{S}^{+*}}{t^*}. \tag{8.12}$$

8.2.5 An Illustrative Example of SBM Models

Table 8.1 exhibits data for Example 1 consisting of six DMUs using two inputs (x_1, x_2) to produce two outputs (y_1, y_2) where (I) and (O) indicate input and output, respectively.

We report the results of the SBM models along with that of the CCR model in Table 8.2. The input-oriented CCR model is described as follows:

[CCR-I]

$$\begin{aligned}
&\theta^* = \min_{\theta, \boldsymbol{\lambda}, \mathbf{s}^-, \mathbf{s}^+} \theta, \\
&\text{subject to} \\
&\theta x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- \quad (i = 1, \ldots, m), \\
&y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ \quad (r = 1, \ldots, s), \\
&\lambda_j \geq 0 \ (\forall j), \quad s_i^- \geq 0 \ (\forall i), \quad s_r^+ \geq 0 \ (\forall r).
\end{aligned} \tag{8.13}$$

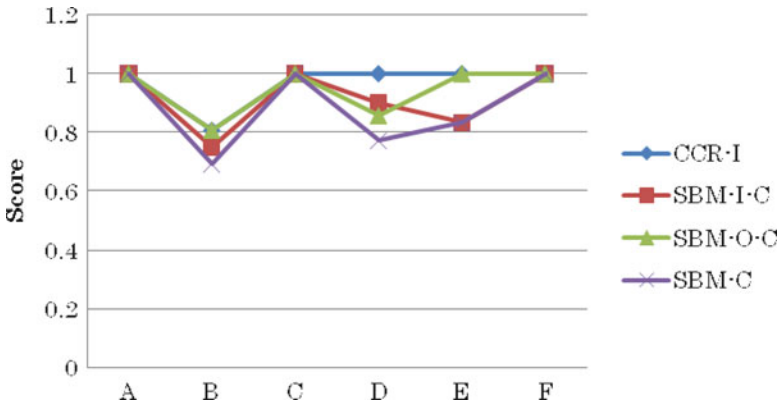


Fig. 8.1 Comparisons of scores for Example 1

Table 8.3 Optimal slacks for CCR-I and SBM-I-C

	CCR-I					SBM-I-C				
DMU	θ^*	s_1^{-*}	s_2^{-*}	s_1^{+*}	s_2^{+*}	ρ_I^*	s_1^{-*}	s_2^{-*}	s_1^{+*}	s_2^{+*}
A	1	0	0	0	0	1	0	0	0	0
B	0.8085106	0	0	0	0	0.75	0	3	0	1
C	1	0	0	0	0	1	0	0	0	0
D	1	4	0	0	2	0.9	4	0	0	2
E	1	16	0	0	0	0.8333333	16	0	0	0
F	1	0	0	0	0	1	0	0	0	0

The CCR-I model found five DMUs out of six efficient. This caused by the radial nature of the model although slacks remain in some of them. However, the SBM models deal with slacks directly and found DMUs *D* and *E* inefficient. In the SBM-O-C model, DMU *E* judged to be efficient, since this DMU has no output slacks. Figure 8.1 compares the scores graphically.

Table 8.3 exhibits the optimal slacks for the CCR-I and the SBM-I-C models. DMUs *D* and *E* have positive slacks in some input and/or output. The CCR model does not account them in efficiency measure. However, the SBM-I-C model accounts the input slacks into efficiency measurement, and DMUs *D* and *E* are judged inefficient.

8.2.6 The Dual Program of the SBM Model

The dual program of [SBM-C-LP] can be expressed as follows, with the dual variables $\mathbf{v} \in R^m$ and $\mathbf{u} \in R^s$:

[SBM-C-LP-Dual]

$$
\begin{aligned}
&\max_{\xi,\mathbf{v},\mathbf{u}} \xi, \\
&\text{subject to} \\
&\xi + \mathbf{v}\mathbf{x}_o - \mathbf{u}\mathbf{y}_o = 1, \\
&-\mathbf{v}\mathbf{X} + \mathbf{u}\mathbf{Y} \leq \mathbf{0}, \\
&\mathbf{v} \geq \frac{1}{m}\left[\frac{1}{\mathbf{x}_o}\right], \\
&\mathbf{u} \geq \frac{\xi}{s}\left[\frac{1}{\mathbf{y}_o}\right],
\end{aligned} \tag{8.14}
$$

where the notation $[1/\mathbf{x}_o]$ designates the row vector $(1/x_{1o}, 1/x_{2o}, \ldots, 1/x_{mo})$. By eliminating ξ from the above program, we have the following equivalent program.

$$
\begin{aligned}
&\max_{\mathbf{v},\mathbf{u}} \mathbf{u}\mathbf{y}_o - \mathbf{v}\mathbf{x}_o, \\
&\text{subject to} \\
&-\mathbf{v}\mathbf{X} + \mathbf{u}\mathbf{Y} \leq \mathbf{0}, \\
&\mathbf{v} \geq \frac{1}{m}\left[\frac{1}{\mathbf{x}_o}\right], \\
&\mathbf{u} \geq \frac{1 - \mathbf{v}\mathbf{x}_o + \mathbf{u}\mathbf{y}_o}{s}\left[\frac{1}{\mathbf{y}_o}\right].
\end{aligned} \tag{8.15}
$$

The dual variables $\mathbf{v} \in R^m$ and $\mathbf{u} \in R^s$ can be interpreted as the virtual costs and prices of input and output items, respectively. The dual program aims to find the optimal virtual costs and prices for DMU $(\mathbf{x}_o, \mathbf{y}_o)$ so that the profit $\mathbf{u}\mathbf{y}_j - \mathbf{v}\mathbf{x}_j$ does not exceed zero for any DMU (including $(\mathbf{x}_o, \mathbf{y}_o)$), and maximize the profit $\mathbf{u}\mathbf{y}_o - \mathbf{v}\mathbf{x}_o$ for the target DMU $(\mathbf{x}_o, \mathbf{y}_o)$. Apparently, the optimal profit is at best zero and hence $\xi^* = 1$ for the SBM-C efficient DMUs.

8.3 Extensions of the SBM Model

In this section, we extend the SBM model to the VRS environment, and introduce the weighted-SBM and the super-SBM models.

8.3.1 Variable Returns-to-Scale Model

All models can be adjusted to the VRS environment by adding the constraint $\mathbf{e}\boldsymbol{\lambda} = \sum_{j=1}^{n} \lambda_j = 1$ where $\mathbf{e}$ denotes a row vector in which all elements are equal to one. Thus, the production possibility set is modified to

$$P_{\mathrm{VRS}} = \left\{ (\mathbf{x}, \mathbf{y}) \middle| x_i \geq \sum_{j=1}^{n} x_{ij} (\forall i), \quad 0 \leq y_r \leq \sum_{j=1}^{n} y_{rj} (\forall r), \quad \mathbf{e}\boldsymbol{\lambda} = 1, \boldsymbol{\lambda} \geq \mathbf{0} \right\}. \tag{8.16}$$

For example, input-oriented SBM under VRS can be defined as follows:

[SBM-I-V] (Input-oriented SBM under variable returns-to-scale assumption)

$$\begin{aligned}
&\rho_I^* = \min_{\boldsymbol{\lambda}, \mathbf{s}^-, \mathbf{s}^+} 1 - \frac{1}{m} \sum_{i=1}^{m} \frac{s_i^-}{x_{io}}, \\
&\text{subject to} \\
&x_{io} = \sum_{j=1}^{n} x_{ij} \lambda_j + s_i^- \quad (i = 1, \ldots, m), \\
&y_{ro} = \sum_{j=1}^{n} y_{rj} \lambda_j - s_r^+ \quad (r = 1, \ldots, s), \\
&\sum\nolimits_{j=1}^{n} \lambda_j = 1, \\
&\lambda_j \geq 0 \ (\forall j), \quad s_i^- \geq 0 \ (\forall i), \quad s_r^+ \geq 0 \ (\forall r).
\end{aligned} \tag{8.17}$$

Similarly, we can define [SBM-O-V] and [SBM-V] models.

8.3.2 Weighted-SBM Model

We can assign weight to input and output slacks in the objective function of (8.10) corresponding to the relative importance of items as follows:

[Weighted-SBM-C]

$$\begin{aligned}
&\rho_{IO}^* = \min_{\boldsymbol{\lambda}, \mathbf{s}^-, \mathbf{s}^+} \frac{1 - (1/m) \sum_{i=1}^{m} (w_i^- s_i^- / x_{io})}{1 + (1/s) \sum_{r=1}^{s} (w_r^+ s_r^+ / y_{ro})}, \\
&\text{subject to} \\
&x_{io} = \sum_{j=1}^{n} x_{ij} \lambda_j + s_i^- \quad (i = 1, \ldots, m), \\
&y_{ro} = \sum_{j=1}^{n} y_{rj} \lambda_j - s_r^+ \quad (r = 1, \ldots, s), \\
&\lambda_j \geq 0 \ (\forall j), \quad s_i^- \geq 0 \ (\forall i), \quad s_r^+ \geq 0 (\forall r),
\end{aligned} \tag{8.18}$$

with $\sum_{i=1}^{m} w_i^- = m$ and $\sum_{r=1}^{m} w_r^+ = s$.The weights should reflect the intentions of decision-makers. We can define the input- (output-) oriented Weighted-SBM models by neglecting the denominator (numerator) of the objective function in (8.18).

8.3.3 Super-SBM Model

In order to rank the SBM-efficient DMUs, we can use the Super-SBM models. Suppose that $\mathrm{DMU}_o = (\mathbf{x}_o, \mathbf{y}_o)$ is SBM-efficient, i.e., $\rho^*_{\mathrm{IO}} = 1$, $\mathbf{s}^- = \mathbf{0}$, and $\mathbf{s}^{+*} = \mathbf{0}$. We define the super-efficiency of $(\mathbf{x}_o, \mathbf{y}_o)$ as the optimal objective function value δ^* for the following program:

[Super-SBM-C]

$$
\begin{aligned}
&\delta^* = \min_{\overline{\mathbf{x}},\overline{\mathbf{y}},\boldsymbol{\lambda}} \frac{(1/m)\sum_{i=1}^{m}(\overline{x}_i/x_{io})}{(1/s)\sum_{r=1}^{s}(\overline{y}_r/y_{ro})},\\
&\text{subject to}\\
&\overline{x}_i \geq \sum_{j=1,j\neq o}^{n} x_{ij}\lambda_j \quad (i = 1,\ldots,m),\\
&\overline{y}_r \leq \sum_{j=1,j\neq o}^{n} y_{rj}\lambda_j \quad (r = 1,\ldots,s),\\
&\overline{\mathbf{x}} \geq \mathbf{x}_o, \quad \overline{\mathbf{y}} \leq \mathbf{y}_o,\\
&\overline{\mathbf{y}} \geq \mathbf{0}, \quad \boldsymbol{\lambda} \geq \mathbf{0}.
\end{aligned}
\tag{8.19}
$$

We can interpret this index as follows. The numerator is a weighted l_1 distance from $\mathbf{x}_o$ to $\overline{\mathbf{x}}(\geq \mathbf{x}_o)$, and hence it expresses an average expansion rate of $\mathbf{x}_o$ to $\overline{\mathbf{x}}(\geq \mathbf{x}_o)$. The denominator is a weighted l_1 distance from $\mathbf{y}_o$ to $\overline{\mathbf{y}}(\leq \mathbf{y}_o)$, and hence it is an average reduction rate of $\mathbf{y}_o$ to $\overline{\mathbf{y}}(\leq \mathbf{y}_o)$. The smaller the denominator is, the farther $\mathbf{y}_o$ is positioned relative to $\overline{\mathbf{y}}$. Its inverse can be interpreted as an index of the distance from $\mathbf{y}_o$ to $\overline{\mathbf{y}}$. Therefore, the objective function in (8.19) is a product of two indices: one, the distance in the input space, and the other in the output space. Both indices are dimensionless. This fractional program can be solved by transforming it into a linear program (see Tone 2002 or Cooper et al. 2007). Dealing with the numerator or the denominator of the objective function, we have input- or output-oriented super-SBM as follows:

[Super-SBM-I-C]

$$
\begin{aligned}
&\delta^*_I = \min_{\overline{\mathbf{x}},\overline{\mathbf{y}},\boldsymbol{\lambda}} \frac{1}{m}\sum_{i=1}^{m}\frac{\overline{x}_i}{x_{io}},\\
&\text{subject to}\\
&\overline{x}_i \geq \sum_{j=1,j\neq o}^{n} x_{ij}\lambda_j \quad (i = 1,\ldots,m),\\
&\overline{y}_r \leq \sum_{j=1,j\neq o}^{n} y_{rj}\lambda_j \quad (r = 1,\ldots,s),\\
&\overline{\mathbf{x}} \geq \mathbf{x}_o, \quad \overline{\mathbf{y}} \leq \mathbf{y}_o,\\
&\overline{\mathbf{y}} \geq \mathbf{0}, \quad \boldsymbol{\lambda} \geq \mathbf{0}.
\end{aligned}
\tag{8.20}
$$

Table 8.4 Super-SBM scores

	Super-SBM-C		Super-SBM-I-C		Super-SBM-O-C	
DMU	Score	Rank	Score	Rank	Score	Rank
A	1.3	1	1.375	1	1.3333333	1
B	0.6923077	6	0.75	6	0.8067227	6
C	1.1428571	2	1.2083333	2	1.1428571	2
D	0.7714286	5	0.9	4	0.8571429	5
E	0.8333333	4	0.8333333	5	1	4
F	1.0909091	3	1.1	3	1.0909091	3

[Super-SBM-O-C]

$$\begin{aligned}
&\delta_O^* = \min_{\overline{\mathbf{x}},\overline{\mathbf{y}},\boldsymbol{\lambda}} \frac{1}{(1/s)\sum_{r=1}^{s}(\overline{y}_r/y_{ro})},\\
&\text{subject to}\\
&\overline{x}_i \geq \sum_{j=1,j\neq o}^{n} x_{ij}\lambda_j \quad (i=1,\ldots,m),\\
&\overline{y}_r \leq \sum_{j=1,j\neq o}^{n} y_{rj}\lambda_j \quad (r=1,\ldots,s),\\
&\overline{\mathbf{x}} \geq \mathbf{x}_o, \quad \overline{\mathbf{y}} \leq \mathbf{y}_o,\\
&\overline{\mathbf{y}} \geq \mathbf{0}, \quad \boldsymbol{\lambda} \geq \mathbf{0}.
\end{aligned} \tag{8.21}$$

8.3.4 An Illustrative Example of Super-SBM Models

Table 8.4 displays the super-efficiency scores for efficient DMUs *A*, *C*, and *F* in Example 1. In any model, DMU *A* is ranked the best followed by *C* (second), and *F* (third).

8.4 Further Extensions

We further extend the SBM model to the nonpositive dataset case. Then, we introduce a variation of the SBM model and finally a composite model which connects the radial and nonradial DEA models.

8.4.1 Dealing with Nonpositive Data in the SBM Models

So far, we have dealt with positive dataset ($\mathbf{X}$, $\mathbf{Y}$). However, occasionally, we encounter nonpositive data. In many DEA models, it is a crucial subject how to deal

with negative data in the evaluation of efficiency. Especially, in the SBM models, nonpositive data effect the scheme critically, since the terms s_i^-/x_{io} and/or s_r^+/y_{ro} lose their original meaning as slack-ratio. If $x_{io} \leq 0$, then we may replace x_{io} by a small positive number and neglect the term s_i^-/x_{io} in the objective function. By the nature of things, inputs should be positive, since it is unreasonable that negative inputs produce positive outputs. However, negative output data should have their duly role in measuring efficiency. As an output, a large deficit (loss) is worse than a small one. In this section, we propose a scheme for resolving this problem. As demonstrated in (8.10), the SBM-C model has the objective function as described below.

$$\min \rho = \frac{1 - (1/m)\sum_{i=1}^{m} s_i^-/x_{io}}{1 + (1/s)\sum_{r=1}^{s} s_r^+/y_{ro}}. \tag{8.22}$$

Let us suppose $y_{ro} \leq 0$. We define $\overline{y}_r^+$ and $\underline{y}_r^+$ by

$$\begin{aligned} \overline{y}_r^+ &= \max_{j=1,\ldots,n} \{y_{rj} | y_{rj} > 0\}, \\ \underline{y}_r^+ &= \min_{j=1,\ldots,n} \{y_{rj} | y_{rj} > 0\}. \end{aligned} \tag{8.23}$$

If the output r has no positive elements, then we define $\overline{y}_r^+ = \underline{y}_r^+ = 1$. We replace the term s_r^+/y_{ro} in the objective function in the following way. (Notice that we never change the value y_{ro} in the constraints.)

1. If $\overline{y}_r^+ > \underline{y}_r^+$, we replace the term by

$$\frac{s_r^+}{(\underline{y}_r^+(\overline{y}_r^+ - \underline{y}_r^+)/\overline{y}_r^+ - y_{ro})}. \tag{8.24}$$

2. If $\overline{y}_r^+ = \underline{y}_r^+$, we replace the term by

$$\frac{s_r^+}{((\underline{y}_r^+)^2/B(\overline{y}_r^+ - y_{ro}))}, \tag{8.25}$$

where B is a large positive number, e.g., $B = 1{,}000$.

In any case, the denominator is positive and strictly less than $\underline{y}_r^+$. Furthermore, it is in inverse proportion to the distance $\overline{y}_r^+ - y_{ro}$. Thus, this scheme takes into account the magnitude of the nonpositive output positively. The score obtained is units invariant, i.e., it is independent of the units of measurement used. Table 8.5 exhibits an example. DMUs D, E, F, and G have nonpositive outputs. The score (right) reflects their magnitude. In this case, we have

$$\overline{y}^+ = 3, \quad \underline{y}^+ = 1.$$

Table 8.5 Negative output data and results

DMU	(I) x	(O) y	Score	Rank
A	1	3	1	1
B	1	2	0.6666667	2
C	1	1	0.3333333	3
D	1	0	0.1818182	4
E	1	−1	0.1111111	5
F	1	−2	0.0740741	6
G	1	−3	5.26E − 02	7

Hence, the denominators of (8.24) for the DMUs with nonpositive output value are as follows:

$$D = \frac{1 \times 2}{3 - 0} = 0.6666, \quad E = \frac{1 \times 2}{3 - (-1)} = 0.5, \quad F = \frac{1 \times 2}{3 - (-2)} = 0.4,$$
$$G = \frac{1 \times 2}{3 - (-3)} = 0.3333.$$

The optimal output slacks (against A) are respectively,

$$s_D^{+*} = 3, \quad s_E^{+*} = 4, \quad s_F^{+*} = 5, \quad s_G^{+*} = 6.$$

Using these values, we obtained the score in Table 8.5 for the SBM-V model. DMU A is the reference to all DMUs.

8.4.2 Variations of the SBM Models

Tone (2010) proposed several variations of the SBM model. Here, we introduce one of them. A drawback of the SBM models is that the projected point is far from the observed point (data). This results from the formulation in which the maximization of distance from the reference set R is required. The reference set spans a facet of the production possibility set and the SBM result seeks the furthest point on the facet from the observed point. Thus, we obtain the worst efficiency score for the DMU. Instead, we can find the nearest point on the facet as follows:

$$\rho_{\mathrm{IO}}^{\max} = \max_{\lambda, \mathbf{s}^-, \mathbf{s}^+} \frac{1 - (1/m) \sum_{i=1}^{m} (s_i^- / x_{io})}{1 + (1/s) \sum_{r=1}^{s} (s_r^+ / y_{ro})},$$

subject to

$$x_{io} = \sum_{j \in R} x_{ij} \lambda_j + s_i^- \quad (i = 1, \ldots, m),$$
$$y_{ro} = \sum_{j \in R} y_{rj} \lambda_j - s_r^+ \quad (r = 1, \ldots, s),$$
$$\lambda_j \geq 0 \ (\forall j), \quad s_i^- \geq 0 \ (\forall i), \quad s_r^+ \geq 0 \ (\forall r). \tag{8.26}$$

This variation demands one additional LP solution for each inefficient DMU. However, since the facet defined by R is an instance of facets and there may be other facets of P to be considered in evaluating the maximum efficiency of DMU $(\mathbf{x}_o, \mathbf{y}_o)$, we need to know all facets of P. See Tone (2010) for details of this subject.

8.4.3 A Compromise of Radial and Nonradial Measures of Efficiency

Tone and Tsutsui (2010) introduced a compromise of radial and nonradial measures, called epsilon-based measure of efficiency (EBM), as follows:

[EBM-I-C] (Input-oriented EBM under CRS)

$$\gamma_{\mathrm{I}}^{*} = \min_{\theta,\boldsymbol{\lambda},\mathbf{s}^{-}} \theta - \varepsilon_x \sum_{i=1}^{m} \frac{w_i^{-} s_i^{-}}{x_{io}},$$

subject to

$$\theta \mathbf{x}_o - \mathbf{X}\boldsymbol{\lambda} - \mathbf{s}^{-} = \mathbf{0},$$

$$\mathbf{Y}\boldsymbol{\lambda} \geq \mathbf{y}_o,$$

$$\boldsymbol{\lambda} \geq \mathbf{0}, \quad \mathbf{s}^{-} \geq \mathbf{0}, \tag{8.27}$$

where w_i^{-} is the weight of input i and satisfies $\sum_{i=1}^{m} w_i^{-} = 1$ $(w_i^{-} \geq 0\ \forall i)$, and ε_x is a key parameter which combines the radial θ and the nonradial slacks term.

Proposition 8.6. If we set $\varepsilon_x = 0$, then EBM-I-C reduces to the CCR-I model.

Proposition 8.7. If we set $\theta = 0$ and $\varepsilon_x = 1$, then EBM-I-C reduces to the Weighted-SBM-I-C model.

Thus, EBM-I-C includes the radial CCR and the nonradial SBM as special cases. Since DEA is a data-driven method, ε_x and $\mathbf{w}^{-}$ are desirable to reflect the characteristics of the dataset. They can be determined by using a sort of the principal component analysis (PCA) on the dataset. See Tone and Tsutsui (2010) for details.

8.5 Concluding Remarks

In this chapter, we have introduced the nonradial SBM models and their extensions. The SBM models utilize the amount of slacks to a maximum extent in measuring efficiency. This might be a merit as well as a demerit. The weighted-SBM models serve to make the models more reliable. This corresponds to the assurance region approach in the radial models. Readers can learn more from the cited references.

References

Cooper WW, Seiford LM, Tone K. Data envelopment analysis: a comprehensive text with models, applications, references and DEA-Solver software. 2nd ed. New York: Springer; 2007.

Pastor JT, Ruiz JL, Sirvent I. An enhanced DEA Russell graph efficiency measure. Eur J Oper Res. 1999;115:596–607.

Tone K. A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res. 2001;130:498–509.

Tone K. A slacks-based measure of super-efficiency in data envelopment analysis. Eur J Oper Res. 2002;143:32–41.

Tone K. Variations on the theme of slacks-based measure of efficiency in DEA. Eur J Oper Res. 2010;200:901–7.

Tone K, Tsutsui M. An epsilon-based measure of efficiency in DEA – a third pole of technical efficiency. Eur J Oper Res. 2010;207:1554–63.

Chapter 9
Chance-Constrained DEA

William W. Cooper, Zhimin Huang, and Susan X. Li

Abstract Incorporation of random variations into DEA analysis has received significant attention in recent years. This chapter describes some of these developments and offers examples of possible uses in the area of chance-constrained programming models in DEA.

Keywords Chance-constrained DEA • Stochastic • Joint probability • Congestion • Satisficing

9.1 Introduction

This chapter deals with chance-constrained programming extensions of the usual deterministic DEA formulations. This kind of approach makes it possible to replace deterministic characterizations in DEA, such as "efficient" and "not efficient," with characterizations such as "probably efficient" and "probably not efficient." Indeed, it is possible to go still further into characterizations such as "sufficiently efficient," with associated probabilities of not being correct in making inferences about the performance of a decision-making unit (DMU).

It is also possible to extend the deterministic objectives usually used in DEA with additional alternatives. For instance, one may use the "E-model" of chance-constrained programming to obtain an "expected value" approach. However, thisexpected value objective is not the only possibility. One may also use the "P-model" of chance-constrained programming to obtain the "most probable" occurrences, perhaps in order to determine whether this probability is sufficiently high. Indeed, one can extend this by incorporating constraints (also probabilistic in character) to insure that the resulting solutions are satisfactory.

Z. Huang (✉)
School of Business, Adelphi University, Garden City, NY 11530, USA
e-mail: huang@adelphi.edu

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_9,

These chance-constrained formulations provide new ways to incorporate new concepts into the DEA literature such as the "satisficing concepts" of H. A. Simon (1957). Originally formulated for use in social psychology these satisficing concepts have now spread to other disciplines such as economics and so it is natural to extend them for use in DEA as is done later in this chapter. See also Cooper et al. (1996).

These are the kinds of ideas and extensions that will be covered in this chapter. The purpose of this chapter, however, is to provide a systematic presentation of major developments of chance-constrained DEA models that have appeared in the literature. The results to be covered are fairly recent so that there is not much to report in the way of significant applications. Indeed, the situation is analogous to the state of game theory and DEA combinations as far as actual applications are concerned. See, for instance, Charnes et al. (1989). As we shall see in the course of developing these chance-constrained approaches to DEA, there is more research to be done (e.g., in the way of developing more efficient algorithms) so this chapter is oriented accordingly.

In the next section, which is Sect. 9.2 of this chapter, we present basic concepts of efficiency and efficiency dominance as well as models to implement these concepts. In Sect. 9.3, we provide a detailed introduction to "joint chance constrained" efficiency and mathematical formulas. Potential uses and deterministic equivalents of the models immediately follow. In Sect. 9.4, we utilize "E-model" (expected value) formulations to discuss DEA efficiency and its relationship with "sensitivity analysis" in stochastic situations. In Sect. 9.5, we briefly summarize another type of chance-constrained DEA models, which are referred to as "P-model" formulations in chance-constrained programming. We then use this class of models to incorporate the "satisficing" concepts of H. A. Simon (1957) for use with DEA. Concluding remarks are in Sect. 9.6.

9.2 Efficiency and Efficiency Dominance

We start with some concepts of "efficiency dominance" (Bowlin et al. 1984) for which we introduce the following notation. Let $x_j = (x_{1j}, \ldots, x_{mj})^{\mathrm{T}}$ and $y_j = (y_{1j}, \ldots, y_{rj})^{\mathrm{T}}$ represent input and output vectors, respectively, for jth DMU, $j = 1, \ldots, n$. The superscript T represents transpose. The DMU to be evaluated is designated as $\mathrm{DMU_o}$ and its input–output vector is denoted (x_0, y_0).

Let us consider a discrete production set which consists of only actually observed input–output vectors, (x_j, y_j), $j = 1, \ldots, n$, as follows:

$$T_0 = \left\{ (x_j, y_j) \right\}_{1 \le j \le n}. \tag{9.1}$$

Definition 9.1 (Efficiency Dominance). DMU_j dominates $\mathrm{DMU_o}$ with respect to T_0 if and only if $x_j \le x_{\mathrm{o}}$ and $y_j \ge y_{\mathrm{o}}$ with strict inequality holding for at least one of the components in the input or output vector.

Thus, DMU_o is *not* dominated in its efficiency if and only if there is *no* DMU_j which exhibits a performance that satisfies the above definition.

One approach to implement Definition 9.1 utilizes the additive model with integer constraints as in Bowlin et al. (1984) as follows:

Dominance Model

$$\begin{aligned}
&\text{Max} \sum_{r=1}^{s} s_r^+ + \sum_{i=1}^{m} s_i^- \\
&\text{s.t.} \\
&x_{i0} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-, \quad i = 1, \ldots, m \\
&y_{r0} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+, \quad r = 1, \ldots, s \\
&1 = \sum_{j=1}^{n} \lambda_j \\
&\lambda_j \in \{0, 1\}, s_i^-, s_r^+ \geq 0, \quad j = 1, \ldots, n; \quad i = 1, \ldots, m; \quad r = 1, \ldots, s.
\end{aligned} \tag{9.2}$$

Solutions with any slacks at nonzero values show the sources and amounts of inefficiency for DMU_o relative to the DMU_j for which $\lambda_j = 1$. The maximization of the slacks in the objective ensures that a "most dominant" DMU_j will be designated for effecting the evaluations. Thus, if the slacks are all zero in a solution to (9.2) then there is *no* DMU_j that dominates DMU_o in the sense of Definition 9.1, above, and this in turn implies that DMU_o operated efficiently relative to all of the other DMU_j.

We can bring this all together by assuming that DMU_k is found to be "most dominant" and then writing the solution to (9.2) as follows:

$$x_{i0} - s_i^{+^*} = \sum_{j=1}^{n} x_{ij}\lambda_j = x_{ik}, \quad i = 1, \ldots, m, \tag{9.3}$$

$$y_{r0} + s_r^{-^*} = \sum_{j=1}^{n} y_{rj}\lambda_j = y_{rk}, \quad r = 1, \ldots, s. \tag{9.4}$$

Using "*" to represent an optimal solution we have let $\lambda_k^* = 1$ with all other $\lambda_j^* = 0$ reflect the fact that an optimal solution to (9.2) has designated DMU_k as "most dominant." DMU_k will dominate DMU_o in the efficiency of its observed performance, however, if and only if *any* $s_i^{+^*}$, $s_r^{-^*}$ is not zero. Conversely, DMU_o is not dominated by *any* DMU_j if and only if all $s_i^{+^*} = s_r^{-^*} = 0$ in an optimal solution for (9.2).

Let us generalize this efficiency dominance to a continuous production possibility set T as follows:

$$T = \left\{ (x, y) : x = \sum_{j=1}^{n} x_j \lambda_j, y = \sum_{j=1}^{n} y_j \lambda_j, \sum_{j=1}^{n} \lambda_j = 1, \quad \lambda_j \geq 0, \quad \forall j \right\} \tag{9.5}$$

Definition 9.2 (General Efficiency Dominance). Let $(x', y') \in T$ and $(x'', y'') \in T$. We say that (x', y') dominates (x'', y'') with respect to the production possibility set T if and only if $x' \leq x''$ and $y' \geq y''$ with strict inequality holding for at least one of the components in the input or the output vector.

Thus, a point in T is not dominated if and only if there is no other point in T which satisfies the definition. This leads to the following definition of efficiency.

Definition 9.3 (Efficiency). DMU_o is efficient with respect to T if and only if there is no $(x, y) \in T$ such that (x_0, y_0) is dominated by (x, y).

A variety of mathematical models are available to implement Definition 9.3. Two models, which are typically employed in the DEA literature are the BCC (Banker et al. 1984) and the Additive model (Charnes et al. 1985). Here, we treat only BCC models but the results also hold for Additive models – and also other DEA models with their associated production possibility sets. Let us consider the following BCC model:

BCC Model

$$\text{Max } \varphi + \varepsilon \left(\sum_{r=1}^{s} s_r^+ + \sum_{i=1}^{m} s_i^- \right)$$

s.t.

$$\varphi y_{r0} - \sum_{j=1}^{n} y_{rj} \lambda_j + s_r^+ = 0, \quad r = 1, \ldots, s,$$

$$\sum_{j=1}^{n} x_{ij} \lambda_j + s_i^- = x_{i0}, \quad i = 1, \ldots, m,$$

$$\sum_{j=1}^{n} \lambda_j = 1,$$

$$\lambda_j \geq 0, s_r^+ \geq 0, s_i^- \geq 0, \quad j = 1, \ldots, n; \quad i = 1, \ldots, m; \quad r = 1, \ldots, s. \tag{9.6}$$

Definition 9.4 (BCC Model). DMU_o is DEA efficient with respect to T if and only if the following two conditions are both satisfied in model (9.6)

(a) $\varphi^* = 1$,
(b) $s_r^{+*} = s_i^{-*} = 0, \quad \forall i, r$,

where "*" represents an optimum.

As in Charnes et al. (1986, 1991), the performances of all DMUs can be partitioned into the following four classes:

$$E, E', F, N,$$

where E is a set of efficient DMUs which are also extreme points and E' is a set of efficient DMUs which are not extreme points. F is a set of points which are on a part of the frontier that is not efficient. Finally, N consists of all points which are not on a frontier and hence are inefficient.

9.3 Stochastic Dominance and Joint Chance Constrained Efficiency

We follow the notation conventions in Cooper et al. (1996, 1998) with $\tilde{x}_j = (\tilde{x}_{1j}, \ldots, \tilde{x}_{mj})^{\mathrm{T}}$ and $\tilde{y}_j = (\tilde{y}_{1j}, \ldots, \tilde{y}_{rj})^{\mathrm{T}}$ to represent random behavior for the output and input vectors for DMU_j, $j = 1, \ldots n$. For the jth DMU, we also let $x_j = (x_{1j}, \ldots, x_{mj})^{\mathrm{T}}$ and $y_j = (y_{1j}, \ldots, y_{rj})^{\mathrm{T}}$ stand for the expected input and output vector values, respectively. The probability distributions of $\tilde{x}_{ij}$ and $\tilde{y}_{rj}$ will usually be determined by historical data on the inputs and outputs but we may replace some or all of these historically determined probability distributions by theoretical probability distributions, as we shall do when this serves our purposes.

For ease of reference, let

$\tilde{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n)$ be the $(s \times n)$ "output" matrix,
$Y = (y_1, \ldots, y_n)$ be the $(s \times n)$ expected "output" matrix,
${}_k\tilde{y} = (\tilde{y}_{k1}, \ldots, \tilde{y}_{kn})$ be the kth row of the "output" matrix $\tilde{Y}$, $k = 1, \ldots, s$,
${}_k y = (y_{k1}, \ldots, y_{kn})$ be the kth row of the expected "output" matrix Y, $k = 1, \ldots, s$,
$\tilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ be the $(m \times n)$ "input" matrix,
$X = (x_1, \ldots, x_n)$ be the $(m \times n)$ expected "input" matrix,
${}_i\tilde{x} = (\tilde{x}_{i1}, \ldots, \tilde{x}_{in})$ be the ith row of the "input" matrix $\tilde{X}$, $i = 1, \ldots, m$,
${}_i x = (x_{i1}, \ldots, x_{in})$ be the ith row of the expected "input" matrix X, $i = 1, \ldots, m$.

Using this notation we can extend our characterizations to "stochastic efficiency dominance" which, for any DMU_o, can be obtained from the *joint* probabilistic comparisons of its outputs and inputs with every other observed DMU. Thus, informally, if $\tilde{y}_0$ and $\tilde{x}_0$ are the output and input vectors of the DMU_o to be tested relative to all DMU_j, $j = 1, \ldots, n$, then we will say that DMU_o is stochastically not dominated in its efficiency if it is stochastically impossible to augment any output without increasing any input and without decreasing any other output, or if it is stochastically impossible to decrease any input without augmenting any other input and without decreasing any output. This is intended as a stochastic generalization of

efficiency dominance as defined in the preceding section. It is also an adaptation of Pareto-Koopmans efficiency to stochastic situations with the discrete production set $\tilde{T}_0 = \left\{(\tilde{x}_j, \tilde{y}_j)\right\}_{1\le j\le n}$. This direct generalization to stochastic situations could be very restricted because of random variations in inputs and outputs. Therefore, we could incorporate a tolerance or risk level to the definition. For a given scalar $\alpha(0 \le \alpha < 1)$, $\mathrm{DMU_o}$ is not stochastically dominated in its efficiency if and only if there is a joint probability less than or equal to α that some other observed DMU displays efficiency dominance relative to $\mathrm{DMU_o}$. Formally (Cooper et al. 1998),

Definition 9.5. $\mathrm{DMU_o}$ is not stochastically dominated in its efficiency with respect to $\tilde{T}_0$ if and only if for all λ satisfying $e^T\lambda = 1$, $\lambda_j \in \{0, 1\}$ – i.e., the components of λ are bivalent – we have

$$P\left\{\bigcap_{i=1}^{m}\left({}_i\tilde{x}\lambda \le \tilde{x}_{i0}\right)\bigcap_{r=1}^{s}\left({}_r\tilde{y}\lambda \ge \tilde{y}_{r0}\right)\right\}$$
$$= P\left\{\sum_{j=1}^{n}\lambda_j\tilde{x}_{ij} \le \tilde{x}_{i0}, \sum_{j=1}^{n}\lambda_j\tilde{y}_{rj} \ge \tilde{y}_{r0}, \quad i = 1,\ldots,m, \quad r = 1,\ldots,s\right\} \le \alpha. \quad (9.7)$$

Note that our definition can be applied to any probability distribution of inputs and outputs for the DMUs to be considered. Also note that if $\tilde{y}_j$ and $\tilde{x}_j$ follow a continuous joint probability distribution, the requirement of "at least one strict inequality" in the above definition is not necessary.

The model in (9.2) is deterministic, so we now provide a stochastic alternative via the following formulation,

$$\text{Max } P\left\{\bigcap_{i=1}^{m}\left({}_i\tilde{x}\lambda \le \tilde{x}_{i0}\right)\bigcap_{r=1}^{s}\left({}_r\tilde{y}\lambda \ge \tilde{y}_{r0}\right)\right\} = \beta. \quad (9.8)$$

Now let $\lambda_k = 1$ be a maximal choice satisfying $e^T\lambda = 1$, $\lambda_j \in \{0, 1\}$, $\forall j$. $\mathrm{DMU_o}$ is then to be regarded as dominated stochastically if and only if $\beta > \alpha$ – in which event DMU_k is designated as the DMU which has the highest probability of dominating $\mathrm{DMU_o}$ in the efficiency of its performance.

We illustrate from a simple example involving four DMUs using one input and one output with known uniform distributions which are also assumed to be independent in order to show more concretely what our definition of stochastic efficiency dominance means. We represent a uniformly distributed random variable $\tilde{z}$ over an interval $a \le z \le b$, by $\tilde{z} \approx \mathrm{Uni[a, b]}$. Hence we write

$$\tilde{y}_1 \approx \text{Uni } [1.5, 2.5], \quad \tilde{y}_2 \approx \text{Uni } [3, 4], \quad \tilde{y}_3 \approx \text{Uni } [4, 5], \quad \tilde{y}_4 \approx \text{Uni } [2.2, 3.2],$$

$$\tilde{x}_1 \approx \text{Uni}[0.5, 1.5], \quad \tilde{x}_2 \approx \text{Uni}[2, 3], \quad \tilde{x}_3 \approx \text{Uni}[4.5, 5.5], \quad \tilde{x}_4 \approx \text{Uni}[3.5, 4.5],$$

to mean that these variables are each uniformly distributed as indicated.

For a given α between 0 and 0.5, suppose it is desired to determine whether DMU_1 is dominated stochastically by any other DMU_j. Substitution in (9.7) produces

$$P\{\lambda_1\tilde{x}_1 + \lambda_2\tilde{x}_2 + \lambda_3\tilde{x}_3 + \lambda_4\tilde{x}_4 \le \tilde{x}_1,\ \lambda_1\tilde{y}_1 + \lambda_2\tilde{y}_2 + \lambda_3\tilde{y}_3 + \lambda_4\tilde{y}_4 \ge \tilde{y}_1\}$$
$$= P\{\tilde{x}_2 \le \tilde{x}_1,\ \tilde{y}_2 \ge \tilde{y}_1\},$$

when we set $\lambda_2 = 1$ and all other $\lambda_j = 0$. The domain of $(\tilde{x}_1, \tilde{x}_2)$ is $[0.5, 1.5] \times [2, 3]$. This does not have any overlap with the area of $\{(x_1, x_2) : x_2 \le x_1\}$, so we have

$$P\{\tilde{x}_2 \le \tilde{x}_1\} = \iint\limits_{x_2 \le x_1} f_1(x_1) f_2(x_2) \mathrm{d}x_1\, \mathrm{d}x_2 = 0.$$

Turning to the outputs, the comparison with DMU_2 in the domain of $(\tilde{y}_1, \tilde{y}_2)$ is $[1.5, 2.5] \times [3, 4]$. All values of y_2 exceed every possible value of y_1. Hence, we have the domain of $(\tilde{y}_1, \tilde{y}_2)$ contained in the area of $\{(y_1, y_2) : y_1 \le y_2\}$ and therefore

$$P\{\tilde{y}_2 \ge \tilde{y}_1\} = \iint\limits_{y_2 \ge y_1} g_1(y_1) g_2(y_2) \mathrm{d}y_1\, \mathrm{d}y_2 = 1.$$

The distributions in our example are all independent. Because of the independence of $\tilde{x}_1, \tilde{x}_2, \tilde{y}_1, \tilde{y}_2$, the joint probability reduces to the product of the probabilities and we have

$$P\{\tilde{x}_2 \le \tilde{x}_1,\ \tilde{y}_2 \ge \tilde{y}_1\} = P\{\tilde{x}_2 \le \tilde{x}_1\} P\{\tilde{y}_2 \ge \tilde{y}_1\} = 0.$$

Therefore, DMU_1 is not dominated stochastically in its efficiency by DMU_2.

Applying the same procedures to compare DMU_1 with DMU_3 and DMU_4, respectively, we also have

$$P\{\tilde{x}_3 \le \tilde{x}_1,\ \tilde{y}_3 \ge \tilde{y}_1\} = 0,$$

$$P\{\tilde{x}_4 \le \tilde{x}_1,\ \tilde{y}_4 \ge \tilde{y}_1\} = 0.$$

Therefore, by Definition 9.5, DMU_1 is *not* dominated stochastically in its efficiency.

It is easy to check that DMU_2 and DMU_3 are also stochastically not dominated in the efficiency of their performances. However, DMU_4 is found to be stochastically dominated in its efficiency because there is a very high probability (98%) that the output of DMU_2 will exceed the output of DMU_4 *and* the probability is unity that the input of DMU_2 will be less than that of DMU_4.

9.3.1 Potential Uses

As described in Sinha (1996), competition in high tech industries can be fast and fierce. Merchant semiconductor manufacturers, for example, face a constant array of new products and processes coming on-stream from competitors and changing demands from users. Merchant semiconductor manufacturers must therefore constantly search for new products of their own to enable them to occupy a new niche, at least for a time, or at least they must develop new processes that will enable them to reduce costs and prices and, in general, they must do both – i.e., develop new products and introduce new processes to constantly reduce prices – as a condition of survival. It is extremely important for the management of such a firm to match itself against the *best* of its competitors, and this must generally be done in the presence of uncertainty as to who these "best" competitors will be, or what they will have to offer in the way of product capabilities, prices, and costs. Thus, we may visualize one potential use of our formulations in situations where such a manufacturer is considering a path of future development in order to continue to survive against competitors for whom, at best, he will know only the probability distributions of pertinent inputs and outputs. Indeed, this manufacturer will know his own output and input prospects only probabilistically since he is considering whether to undertake *proposed* developments. The above formulations can help this manufacturer to locate the potential "best competitors" – i.e., those which are likely to be most efficient with input and output mixes similar to those required for the niche being considered.

As another example, we turn to the 5-year "field experiment" conducted by R.L. Clarke (1989) which is summarized in Charnes et al. (1989). The main objective of the study was directed to examining the possible effects of repeated uses of DEA to evaluate performances when used over an extended period of time. This was done in a context that involved evaluating the performance of vehicle maintenance units located at each of several different bases in the US Air Force's Tactical Air Command. Under orders from Central Headquarters the commanders at each base were required to report results each month in a form suited to DEA evaluations and in ways that conformed to the desired study conditions with the understanding that their vehicle maintenance operations were likely to be subject to on-site inspections if their performance was out of line (i.e., less efficient) than the performances of the maintenance units at other bases.

Noting that our development, as given above, admits a use of degenerate distributions, we can visualize a situation in which a base commander knows his own performance and wants to determine the likelihood of an inspection resulting from performances reported to headquarters by other bases. The performances of *other* bases are not known exactly but all commanders have records of past performance from these other bases which can be used to synthesize probability distributions. Given exact knowledge of performance at his own base, it is then a simple matter for each commander to calculate the probability that one or more of the other bases will report a better performance record in one or more of its inputs or outputs. Conversely, we may

think of Central Headquarters as using such probability distributions to help plan its inspection actions.

Other examples could include reanalyses of earlier DEA studies as exemplified by Land et al. (1993) in their chance-constrained programming re-evaluation of the earlier deterministic treatment by Charnes et al. (1981) of the (huge) Program-Follow-Through experiment in public school education conducted by the US Office (now Department) of Education. Such analyses can become quite complex, of course, but others can be treated in very simple and straightforward ways.

The formulations for stochastic efficiency dominance used above are based on the discrete stochastic production set

$$\begin{aligned}\tilde{T}_0 &= \left\{(\tilde{x},\tilde{y}) : \tilde{x} = \tilde{X}\lambda,\ \tilde{y} = \tilde{Y}\lambda,\ e^T\lambda = 1, \quad \lambda_j \in \{0,1\}, \quad j = 1,\ldots,n\right\} \\ &= \left\{(\tilde{x}_j,\tilde{y}_j)\right\}_{1\le j\le n},\end{aligned} \tag{9.9}$$

where e is a $(n \times 1)$ vector with all elements equal to unity. We extend this to the continuous stochastic production set $\tilde{T}$, in which the bivalency conditions on the variables λ_j are relaxed. These variables are now allowed to be continuous so that the required evaluation can be effected in terms of convex combinations of observed DMUs. Therefore, $\tilde{T}$ can be written as

$$\tilde{T} = \left\{(\tilde{x},\tilde{y}) : \tilde{x} = \tilde{X}\lambda,\ \tilde{y} = \tilde{Y}\lambda,\ e^T\lambda = 1, \quad \lambda \ge 0\right\}. \tag{9.10}$$

One of the associated stochastic production possibility sets of $\tilde{T}$ which is also considered here can be defined as

$$\tilde{T}_1 = \left\{(\tilde{x},\tilde{y}) : \tilde{x} = \tilde{X}\lambda + s^+,\ \tilde{y} = \tilde{Y}\lambda - s^-,\ e^T\lambda = 1,\ \lambda \ge 0,\ s^+ \ge 0,\ s^- \ge 0\right\}. \tag{9.11}$$

We now note that this brings us into contact with other parts of the DEA literature. $\tilde{T}$ and $\tilde{T}_1$ are stochastic generalizations of the production possibility sets defined in Charnes et al. (1985) and Banker et al. (1984), respectively, where the "Additive" and "BCC" models of DEA were first introduced into the literature. To see this, let us represent the general stochastic production possibility set as follows:

$$\tilde{T}_2 = \{(\tilde{x},\tilde{y}) : \tilde{y} \text{ can be produced from } \tilde{x}\}.$$

We next postulate the following properties for the production possibility set (Cooper et al. 1998), $\tilde{T}_2$:

Postulate 1. *Convexity*. If $(\tilde{x}_j,\tilde{y}_j) \in \tilde{T}_2$, $j = 1, \ldots, n$, and $\lambda_j \ge 0$ are non-negative scalars such that $e^T\lambda = 1$, then $(\tilde{X}\lambda, \tilde{Y}\lambda) \in \tilde{T}_2$.

Postulate 2. *Inefficiency Postulate*. (a) If $(\tilde{x},\tilde{y}) \in \tilde{T}_2$ and $\tilde{x}^* = \tilde{x} + s^+$ with $s^+ \ge 0$, then $(\tilde{x}^*,\tilde{y}) \in \tilde{T}_2$. (b) If $(\tilde{x},\tilde{y}) \in \tilde{T}_2$ and $\tilde{y}^* = \tilde{y} - s^-$ with $s^- \ge 0$, then $(\tilde{x},\tilde{y}^*) \in \tilde{T}_2$.

Postulate 3. *Minimum Intersection.* $\tilde{T}_2$ is the intersection set of all $\hat{T}$ satisfying Postulates 1 and 2, and subject to the condition that each of the vectors $(\tilde{x}_j, \tilde{y}_j) \in \hat{T}$, $j = 1, \ldots, n$.

$$\tilde{T}_1 = \left\{(\tilde{x}, \tilde{y}) : \tilde{x} = \tilde{X}\lambda + s^+, \tilde{y} = \tilde{Y}\lambda - s^-, e^T\lambda = 1, \lambda \geq 0, s^+ \geq 0, s^- \geq 0\right\}$$

is thus a stochastic production possibility set satisfying the above three postulates. Furthermore, if we omit the convexity condition $e^T\lambda = 1$ in $\tilde{T}_1$ the production possibility set becomes a stochastic generalization of the production possibility set for the CCR model introduced in Charnes et al. (1978). Cooper et al. (1998) have shown that both $\tilde{T}$ and $\tilde{T}_1$ have the same efficiency properties. Therefore, here we only discuss some major results on $\tilde{T}$.

We can now replace "stochastic efficiency dominance" with a more general concept of "stochastic efficiency" on $\tilde{T}$.

Definition 9.6. For $0 \leq \alpha < 1$, $(\tilde{x}^*, \tilde{y}^*) \in \tilde{T}$ is "α-stochastically efficient" with respect to $\tilde{T}$ if for any λ satisfying $e^T\lambda = 1$ and $\lambda \geq 0$, we have

$$P\left\{\bigcap_{i=1}^{m}\left({}_i\tilde{x}\lambda \leq \tilde{x}_i^*\right)\bigcap_{r=1}^{s}\left({}_r\tilde{y}\lambda \geq \tilde{y}_r^*\right)\right\} \leq \alpha. \tag{9.12}$$

Definition 9.7. The α-stochastically efficient frontier of $\tilde{T}$ is defined as the set of α-stochastically efficient points for which there exists a λ satisfying $e^T\bar{\lambda} = 1$ with $\bar{\lambda} \geq 0$ such that equality holds in (9.12).

Definition 9.8. DMU_o is α-stochastically efficient if for any λ satisfying $e^T\lambda = 1$ and $\lambda \geq 0$, we have

$$P\left\{\bigcap_{i=1}^{m}\left({}_i\tilde{x}\lambda \leq \tilde{x}_{i0}\right)\bigcap_{r=1}^{s}\left({}_r\tilde{y}\lambda \geq \tilde{y}_{r0}\right)\right\} \leq \alpha. \tag{9.13}$$

The rest of this section is devoted to sharpening our characterizations of α-stochastic efficiencies. Since

$$\left(\bigcap_{i=1}^{m}\left({}_i\tilde{x}\lambda \leq \tilde{x}_{i0}\right)\bigcap_{r=1}^{s}\left({}_r\tilde{y}\lambda \geq \tilde{y}_{r0}\right)\right) \subset \left\{e^T\tilde{X}\lambda - e^T\tilde{Y}\lambda < e^T\tilde{x}_0 - e^T\tilde{y}_0\right\},$$

it follows that

$$P\left(\bigcap_{i=1}^{m}\left({}_i\tilde{x}\lambda \leq \tilde{x}_{i0}\right)\bigcap_{r=1}^{s}\left({}_r\tilde{y}\lambda \geq \tilde{y}_{r0}\right)\right) \leq P\left\{e^T\tilde{X}\lambda - e^T\tilde{Y}\lambda < e^T\tilde{x}_0 - e^T\tilde{y}_0\right\}.$$

Therefore, $P\left\{e^T\tilde{X}\lambda - e^T\tilde{Y}\lambda < e^T\tilde{x}_0 - e^T\tilde{y}_0\right\} \leq \alpha$ is sufficient for DMU_o to be α-stochastically efficient. Next, we let ε be the non-Archimedean positive infinitesimal. The following theorem develops a necessary condition for a DMU to be α-stochastically efficient.

Theorem 9.1 (Cooper et al. 1998). Let DMU_o be α-stochastically efficient. Then for any λ which satisfies

$$P\{{}_i\tilde{x}\lambda < \tilde{x}_{i0}\} \geq 1 - \varepsilon, \quad i = 1, \ldots, m, \tag{9.14}$$

$$P\{{}_k\tilde{y}\lambda > \tilde{y}_{k0}\} \geq 1 - \varepsilon, \quad k = 1, \ldots, s, \tag{9.15}$$

$$e^T\lambda = 1, \quad \lambda \geq 0, \tag{9.16}$$

we have

$$P\left\{e^T\tilde{X}\lambda - e^T\tilde{Y}\lambda < e^T\tilde{x}_0 - e^T\tilde{y}_0\right\} \leq \alpha. \tag{9.17}$$

Theorem 9.1 allows us to develop an extension of (9.2) for use in evaluating stochastic efficiency because it implies that if DMU_o is α-stochastically efficient, then the maximum value of the chance functional $P\left\{e^T(\tilde{X}\lambda - \tilde{x}_0) + e^T(\tilde{y}_0 - \tilde{Y}\lambda) < 0\right\}$, subject to constraints (9.14)–(9.17), is less than or equal to the specified risk level α. It is obvious that if the maximum value of the chance functional $P\left\{e^T(\tilde{X}\lambda - \tilde{x}_0) + e^T(\tilde{y}_0 - \tilde{Y}\lambda) < 0\right\}$ exceeds α, then DMU_o is not α-stochastically efficient. Since ε is a positive non-Archimedean infinitesimal, we call (9.14) and (9.15) "almost 100% confidence" chance constraints. This is reasonable because we are almost 100% sure that for any point in $\tilde{T}$, which satisfies constraints (9.14) and (9.15), its individual inputs and outputs are less than and greater than the corresponding inputs and outputs of DMU_o, respectively. Hence, at least in principle, the determination of α-stochastic efficiency of all DMUs can be characterized by solving a series of almost 100% confidence chance-constrained programming problems.

To represent this explicitly, we introduce the "almost 100% confidence" chance-constrained problem represented in (9.18)–(9.22),

$$\text{Max } P\left\{e^T(\tilde{X}\lambda - \tilde{x}_0) + e^T(\tilde{y}_0 - \tilde{Y}\lambda) < 0\right\}$$

$$\text{s.t.} \tag{9.18}$$

$$P\{{}_i\tilde{x}\lambda < \tilde{x}_{i0}\} \geq 1 - \varepsilon, \quad i = 1, \ldots, m, \tag{9.19}$$

$$P\{{}_k\tilde{y}\lambda > \tilde{y}_{k0}\} \geq 1 - \varepsilon, \quad k = 1, \ldots, s, \tag{9.20}$$

$$e^T\lambda = 1, \quad \lambda \geq 0, \tag{9.21}$$

$$\lambda \geq 0.Z \tag{9.22}$$

From the above discussions we then have the following:

Theorem 9.2 (Cooper et al. 1998). (1) Let DMU_o be α-stochastically efficient. Then the optimal objective value of the "almost 100% confidence" chance-constrained programming problem (9.18)–(9.22) is less than or equal to α. (2) If the optimal objective value of (9.18) exceeds α, then DMU_o is not stochastically efficient.

We now undertake further developments which depend on explicit assumptions for the types of probability distributions to be used. A simple, frequently used approach, is to suppose that $\tilde{x}_{ij}$, $\tilde{y}_{kj}$ follow a multivariate normal distribution with means and a covariance matrix as follows:

$$E(\tilde{x}_{ij}) = x_{ij}, \tag{9.23}$$

$$E(\tilde{y}_{kj}) = y_{kj}, \tag{9.24}$$

$$\Delta = \begin{pmatrix} \left(\Delta_{ij}^{II}\right)_{m\times m} & \left(\Delta_{ik}^{IO}\right)_{m\times s} \\ \left(\Delta_{kj}^{OI}\right)_{s\times m} & \left(\Delta_{ij}^{OO}\right)_{s\times s} \end{pmatrix}, \tag{9.25}$$

where

$$\Delta_{ij}^{II} = \left(\mathrm{Cov}(\tilde{x}_{iq}, \tilde{x}_{jr})\right)_{n\times n}, \quad 1 \le i,j \le m, \tag{9.26}$$

$$\Delta_{ik}^{IO} = \left(\mathrm{Cov}(\tilde{x}_{iq}, \tilde{y}_{kr})\right)_{n\times n}, \quad 1 \le i \le m, \quad 1 \le k \le s, \tag{9.27}$$

$$\Delta_{ij}^{OO} = \left(\mathrm{Cov}(\tilde{y}_{iq}, \tilde{y}_{jr})\right)_{n\times n}, \quad 1 \le i,j \le s, \tag{9.28}$$

$$\Delta_{ik}^{IO} = \Delta_{ki}^{OI}, \quad 1 \le i \le m, \quad 1 \le k \le s. \tag{9.29}$$

In order to simplify our model development, we also introduce new notations as follows:

$$\left(\sigma_i^I(\lambda)\right)^2 = V(\tilde{x}_{i0} - {}_i\tilde{x}\lambda) = \lambda^T(\Delta_{ii}^{II})\lambda - 2\sum_{j=1}^{n}\lambda_j \mathrm{Cov}(\tilde{x}_{ij}, \tilde{x}_{i0}) + V(\tilde{x}_{i0}), \tag{9.30}$$

$$\left(\sigma_k^O(\lambda)\right)^2 = V(\tilde{y}_{k0} - {}_k\tilde{y}\lambda) = \lambda^T(\Delta_{kk}^{OO})\lambda - 2\sum_{i=1}^{n}\lambda_j \mathrm{Cov}(\tilde{y}_{ki}, \tilde{y}_{k0}) + V(\tilde{y}_{k0}), \tag{9.31}$$

$$\begin{aligned}
(\sigma(\lambda))^2 &= V(e^T(\tilde{X}\lambda - \tilde{x}_0) + e^T(\tilde{y}_0 - \tilde{Y}\lambda)) \\
&= \lambda^T\left[\sum_{i=1}^{m}\sum_{j=1}^{m}\Delta_{ij}^{II} + \sum_{k=1}^{s}\sum_{j=1}^{s}\Delta_{kj}^{OO} - 2\sum_{i=1}^{m}\sum_{k=1}^{s}\Delta_{ik}^{IO}\right]\lambda \\
&\quad + 2\sum_{p=1}^{n}\lambda_p\left[\sum_{i=1}^{m}\sum_{k=1}^{s}\mathrm{Cov}(\tilde{x}_{i0}, \tilde{y}_{kp}) + \sum_{i=1}^{m}\sum_{k=1}^{s}\mathrm{Cov}(\tilde{x}_{ip}, \tilde{y}_{k0})\right. \\
&\quad \left. - \sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{Cov}(\tilde{x}_{i0}, \tilde{x}_{jp}) - \sum_{k=1}^{s}\sum_{i=1}^{s}\mathrm{Cov}(\tilde{y}_{kp}, \tilde{y}_{i0})\right] \\
&\quad + \left[\sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{Cov}(\tilde{x}_{i0}, \tilde{x}_{j0}) - 2\sum_{i=1}^{m}\sum_{k=1}^{s}\mathrm{Cov}(\tilde{x}_{i0}, \tilde{y}_{k0})\right. \\
&\quad \left. + \sum_{k=1}^{s}\sum_{j=1}^{s}\mathrm{Cov}(\tilde{y}_{k0}, \tilde{y}_{j0})\right].
\end{aligned} \tag{9.32}$$

The assumption of multivariate normality implies that the production possibility set and its efficient frontier will vary randomly in a symmetric manner across DMUs, and in this manner reflects the results of events such as bad weather and poor luck, etc., and it also permits data measurement and other errors to occur symmetrically.

This interpretation is similar to the two-sided error assumptions used by Aigner et al. (1977) for the estimation of single output parametric stochastic frontier production functions. There is another one-sided error in their work, which represents a component that reflects an assumption that each DMU's output must lie on or below the stochastic frontier function if it is to represent "inefficiency." Although we do not consider this one-sided disturbance explicitly in our stochastic DEA model, we do need to note that the structure of our stochastic production possibility set implicitly allows for one-sided disturbances from the efficient frontier due to possible DMU inefficiencies and this is reflected in our chance constraints being oriented in the direction where inefficiencies might occur. We restrict our consideration to the class of "zero-order decision rules" in chance-constrained programming to achieve a deterministic equivalent for the problem (9.18)–(9.22) as follows:

$$\begin{aligned} &\text{Min } e^T(X\lambda - x_0) + e^T(y_0 - Y\lambda) + \sigma(\lambda)\Phi^{-1}(\alpha) \\ &\text{s.t.} \end{aligned} \tag{9.33}$$

$$y_{k0} \leq {}_k y\lambda + \sigma_k^O(\lambda)\Phi^{-1}(\varepsilon), \quad k = 1, \ldots, s, \tag{9.34}$$

$${}_i x\lambda \leq x_{i0} + \sigma_i^I(\lambda)\Phi^{-1}(\varepsilon), \quad i = 1, \ldots, m, \tag{9.35}$$

$$e^T\lambda = 1, \tag{9.36}$$

$$\lambda \geq 0.Z \tag{9.37}$$

for which we have the following theorem,

Theorem 9.3 (Cooper et al. 1998). (1) Let DMU_o be α-stochastically efficient. Then the optimal objective value of problem (9.33)–(9.37) is greater than or equal to zero. (2) If the optimal objective value of (9.33) is less than zero, then DMU_o is not α-stochastically efficient.

9.4 Stochastic Efficiency in Marginal Chance Constrained Models

Land et al. (1993) introduced a formal "E-model" form of marginal chance-constrained DEA model in CCR (Charnes et al. 1978) form as follows:

$$
\begin{aligned}
&\text{Min } \theta \\
&\text{s.t.} \\
&P\left\{\theta\tilde{x}_{i0} \geq \sum_{j=1}^{n} \tilde{x}_{ij}\lambda\right\} \geq 1-\alpha, \quad i=1,\ldots,m, \\
&P\left\{\sum_{j=1}^{n} \tilde{y}_{rj}\lambda_j \geq \tilde{y}_{r0}\right\} \geq 1-\alpha, \quad r=1,\ldots,s, \\
&\lambda_j \geq 0, \quad j=1,\ldots,n.
\end{aligned} \tag{9.38}
$$

The meaning of the chance constraints is that they should not be violated with probability at most a.

Olesen and Petersen (1995) also utilized marginal chance-constrained programming theory to develop an "E-model" for use in DEA by introducing confidence regions for all DMUs. For given probability level γ,

$$
D_j(\gamma) = \left\{(x,y) : (x^T - E(\tilde{x}_j)^T, y^T - E(\tilde{y}_j)^T) \sum\nolimits_j^{-1} \begin{pmatrix} x - E(\tilde{x}_j) \\ y - E(\tilde{y}_j) \end{pmatrix} \leq c^2\right\}
$$

is called the confidence region of DMU_j, where $\sum_j$ is the covariance matrix of $(\tilde{x}_j, \tilde{y}_j)$, c is determined by $P\left(\chi^2_{(n)} \leq c^2\right) = \gamma$, and $\chi^2_{(n)}$ is the Chi-square random variable with n degrees of freedom. A comparison of these two types of approaches can be found in Olesen (2006).

Letting $\alpha = 1 - \Phi(c)$, the chance-constrained DEA model is

$$
\begin{aligned}
&\text{Max } u^T y_0 - v^T x_0 \\
&\text{s.t.} \\
&P(u^T y_j \leq v^T x_j) \geq 1-\alpha, \quad j=1,\ldots,n, \\
&u \geq \varepsilon e, v \geq \varepsilon e,
\end{aligned} \tag{9.39}
$$

where ε is the non-Archimedean positive infinitesimal defined scalar and e is a vector of ones.

Differences between models in (9.38) and (9.39) are (a) the model in (9.38) generalized the CCR envelopment form to marginal chance-constrained formulations, while the model in (9.39) extended CCR multiplier models to marginal chance-constrained formulations; (b) the scalar α was predetermined directly by the user in model (9.38), but in model (9.39) the scalar α was determined by another scalar γ through confidence regions of DMUs.

We here consider another version of "E-model" form for marginal chance-constrained DEA models (Cooper et al. 2002a, b, 2003), which we will use to extend the concepts of "DEA efficiency" and congestion of BCC models to a chance-constrained programming context,

$$
\begin{aligned}
&\text{Max } \varphi \\
&\text{s.t.} \\
&P\left\{\sum_{j=1}^{n} \tilde{y}_{rj}\lambda_j \geq \varphi\tilde{y}_{r0}\right\} \geq 1-\alpha, \quad r=1,\ldots,s, \\
&P\left\{\sum_{j=1}^{n} \tilde{x}_{ij}\lambda_j \leq +\tilde{x}_{i0}\right\} \geq 1-\alpha, \quad i=1,\ldots,m, \\
&\sum_{j=1}^{n} \lambda_j = 1, \\
&\lambda_j \geq 0, \quad j=1,\ldots,n.
\end{aligned} \tag{9.40}
$$

Here, α is a predetermined number between 0 and 1.

Definition 9.9 (Chance Constrained Efficiency). DMU_o is stochastic efficient if and only if the following two conditions are both satisfied:

1. $\varphi^* = 1$;
2. Slack values are all zero for *all* optimal solutions.

Here (2) refers to *all* alternate optima because the second stage optimization associated with $\varepsilon > 0$ is not used in (9.40).

Since $j = 0$ is one of the n DMU_j, we can always get a solution with $\phi = 1$, $\lambda_0 = 1$, and $\lambda_j = 0$ ($j \neq 0$) and all slacks zero. However, this solution need not be maximal. It follows that a maximum with $\phi^* > 1$ in (9.40) for any sample of $j = 1, \ldots, n$ observations means that the DMU_o being evaluated is not efficient because, to the specified level of probability defined by α, the evidence will then show that all outputs of DMU_o can be increased to $\varphi^*\tilde{y}_{r0} > \tilde{y}_{r0}$, $r = 1, \ldots, s$, by using a convex combination of other DMUs which also satisfy

$$
P\left\{\sum_{j=1}^{n} \tilde{x}_{ij}\lambda_j \leq \tilde{x}_{i0}\right\} \geq 1-\alpha, \quad i=1,\ldots,m. \tag{9.41}
$$

Hence, as required by Definition 9.3, no output or input is worsened by this increase so, in effect, we have added a stochastic element to the deterministic formulations in Definitions 3 and 4.

Now suppose $\zeta_r > 0$ is the "external slack" for the rth output. By "external slack" we refer to slack outside the braces. We can choose the value of this external slack so it satisfies

$$
P\left\{\sum \tilde{y}_{rj}\lambda_j - \varphi\tilde{y}_{r0} \geq 0\right\} = (1-\alpha) + \zeta_r. \tag{9.42}
$$

There must then exist a positive number $s_r^+ > 0$ such that

$$
P\left\{\sum \tilde{y}_{rj}\lambda_j - \varphi\tilde{y}_{r0} \geq s_r^+\right\} = 1+\alpha. \tag{9.43}
$$

This positive value of s_r^+ permits a still further increase in $\tilde{y}_{r0}$ for any set of sample observations without worsening any other input or output. It is easy to see that $\zeta_r = 0$ if and only if $s_r^+ = 0$.

In a similar manner, suppose $\xi_i > 0$ represents "external slack" for the ith input chance constraint. We choose its value to satisfy

$$P\left\{\sum_{j=1}^{n} \tilde{x}_{ij}\lambda_j - \tilde{x}_{i0} \leq 0\right\} = (1-\alpha) + \xi_i. \tag{9.44}$$

There must then exist a positive number $s_i^- > 0$ such that

$$P\left\{\sum_{j=1}^{n} \tilde{x}_{ij}\lambda_j + s_i^- \leq \tilde{x}_{i0}\right\} = 1 - \alpha. \tag{9.45}$$

Such a positive value of s_i^- permits a decrease in $\tilde{x}_{i0}$ for any sample without worsening any other input or output to the indicated probabilities. It is easy to show that $\xi_i = 0$ if and only if $s_i^- = 0$.

We again introduce the non-Archimedean infinitesimal, $\varepsilon > 0$, and extend (9.40) so that stochastic efficiencies and inefficiencies can be characterized by the following model:

$$\text{Max } \varphi + \varepsilon\left(\sum s_r^+ + \sum s_i^-\right)$$

s.t.

$$P\left\{\sum \tilde{y}_{rj}\lambda_j - \varphi\tilde{y}_{r0} \geq s_r^+\right\} = 1 - -\alpha, \quad r = 1, \ldots, s,$$

$$P\left\{\sum_{j=1}^{n} \tilde{x}_{ij}\lambda_j + s_i^- \leq \tilde{x}_{i0}\right\} = 1 - \alpha, \quad i = 1, \ldots, m,$$

$$\sum_{j=1}^{n} \lambda_j = 1,$$

$$\lambda_j \geq 0, s_r^+ \geq 0, s_i^- \geq 0, \quad j = 1, \ldots, n; \quad r = 1, \ldots, s; \quad i = 1, \ldots, m. \tag{9.46}$$

This leads to the following modification of Definition 9.9.

Definition 9.10. DMU_o is stochastic efficient if and only if the following two conditions are both satisfied

1. $\varphi^* = 1$,
2. $s_r^{+*} = 0$, $s_i^{-*} = 0, \forall i, r$.

This definition aligns more closely with Definition 9.4 since the $\varepsilon > 0$ in the objective of (9.46) makes it unnecessary to refer to "*all* optimal solutions," as in Definition 9.9. It differs from Definition 9.4, however, in that it refers to stochastic characterizations. Thus, even when the conditions of Definition 9.10 are satisfied there is a chance (determined by the choice of α) that the thus characterized DMU_o is not efficient.

The stochastic model in (9.46) is evidently a generalization of the BCC model in (9.4). Assume that inputs and outputs are random variables with a multivariate normal distribution and known parameters. The deterministic equivalent for model (9.46) is as follows:

$$
\begin{aligned}
&\text{Max } \varphi + \varepsilon\left(\sum s_r^+ + \sum s_i^-\right)\\
&\text{s.t.}\\
&\varphi y_{rn} - \sum y_{rj}\lambda_j + s_r^+ - \Phi^{-1}(\alpha)\sigma_r^o(\varphi,\lambda) = 0, \quad r = 1,\ldots,s,\\
&\sum x_{ij}\lambda_j + s_i^- - \Phi^{-1}(\alpha)\sigma_j^I(\lambda) = x_{i0}, \quad i = 1,\ldots,m,\\
&\sum \lambda_j = 1,\\
&\lambda_j \geq 0, s_r^+ \geq 0, s_i^- \geq 0, \quad j = 1,\ldots,n; \quad r = 1,\ldots,s; \quad i = 1,\ldots,m,
\end{aligned}
\tag{9.47}
$$

were Φ is the standard normal distribution function and Φ^{-1}, its inverse, is the so-called fractile function. Finally,

$$
\begin{aligned}
\left(\sigma_r^o(\varphi,\lambda)\right)^2 = &\sum_{i\neq 0}\sum_{j\neq 0}\lambda_i\lambda_j\text{Cov}(\tilde{y}_{ri},\tilde{y}_{rj}) + 2(\lambda_0-\varphi)\sum_{i\neq 0}\lambda_i\text{Cov}(\tilde{y}_{ri},\tilde{y}_{r0})\\
&+ (\lambda_0-\varphi)^2\text{Var}(\tilde{y}_{r0})
\end{aligned}
$$

and

$$
\begin{aligned}
\left(\sigma_i^I(\lambda)\right)^2 = &\sum_{j\neq 0}\sum_{k\neq 0}\lambda_j\lambda_k\text{Cov}(\tilde{x}_{ij},\tilde{x}_{ik}) + 2(\lambda_0-1)\sum_{j\neq 0}\lambda_j\text{Cov}(\tilde{x}_{ij},\tilde{x}_{i0})\\
&+ (\lambda_0-1)^2\text{Var}(\tilde{x}_{i0}),
\end{aligned}
$$

where we have separated out the terms for DMU_o because they appear on both sides of the expressions in (9.46). Thus, φ^*, s_r^{+*}, and s_i^{-*} can be determined from (9.47) where the data (means and variances) are all assumed to be known.

Let us simplify our assumptions in a manner that will enable us to relate what we are doing to other areas such as the sensitivity analysis research in DEA that is reported in Cooper et al. (2001). Therefore, assume that only DMU_o has random

variations in its inputs and outputs and they are statistically independent. In this case, model (9.47) can be written in the following simpler form:

$$\text{Max}\varphi + \varepsilon\left(\sum s_r^+ + \sum s_i^-\right)$$

s.t.

$$\varphi y'_{r0} - \sum y'_{rj}\lambda_j + s_r^+ = 0, \quad r = 1, \ldots, s,$$

$$\sum x'_{ij}\lambda_j + s_i^- = x'_{i0}, \quad I = 1, \ldots, m,$$

$$\sum \lambda_j = 1,$$

$$\lambda_j \geq 0, s_r^+ \geq 0, s_i^- \geq 0, \quad j = 1, \ldots, n; \quad r = 1, \ldots, s; \quad i = 1, \ldots, m, \tag{9.48}$$

where

$$y'_{r0} = y_{r0} - \sigma_{r0}^o \Phi^{-1}(\alpha), \quad r = 1, \ldots, s,$$

$$y'_{rj} = y_{rj}, \quad j \neq 0, \quad r = 1, \ldots, s,$$

$$x'_{i0} = x_{i0} + \sigma_{i0}^I \Phi^{-1}(\alpha), \quad i = 1, \ldots, m,$$

$$x'_{ij} = x_{ij}, \quad j \neq 0, \quad i = 1, \ldots, m,$$

$$\sigma_{r0}^o = \sqrt{\text{Var}(y_{r0})},$$

$$\sigma_{r0}^I = \sqrt{\text{Var}(x_{i0})}.$$

Reasons for us to consider random variations only in DMU_o are as follows: First, treating more than one DMU in this manner leads to deterministic equivalents with the more complicated relations that have been discussed in detail in Cooper et al. (2002a, b, 2003). The simpler approach used here allows us to arrive at analytical results and characterizations in a straightforward manner. Second, it opens possible new routes for effecting "sensitivity analyses." We are referring to the "sensitivity analyses" that are to be found in Charnes and Neralic (1990), Charnes et al. (1992), Charnes et al. (1996), Seiford and Zhu (1998). In the terminology of the survey article by Cooper et al. (2001), these sensitivity analyses are directed to analyzing allowable limits of data variations for only one DMU at a time and hence contrast with other approaches to sensitivity analysis in DEA that allow all data for all DMUs to be varied simultaneously until at least one DMU changes its status from efficient to inefficient, or vice versa. These sensitivity analyses are entirely deterministic. Our chance-constrained approach can be implemented by representations that are similar in form to those used in sensitivity analysis but the conceptual meanings are different. A chance-constrained programming problem can be solved by a deterministic equivalent, as we have just shown, but the issue originally addressed in the chance-constrained formulation is different and this introduces elements, such as the risk associated with α, that are nowhere present in these sensitivity analyses.

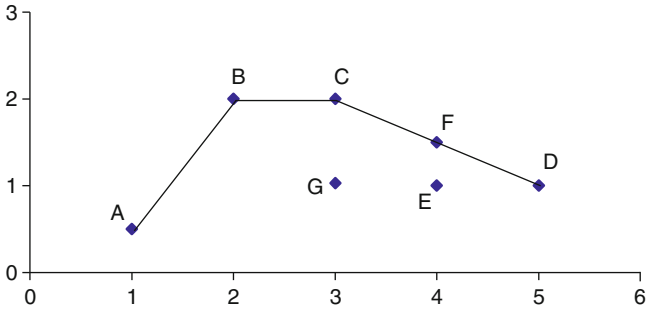


Fig. 9.1 Technical inefficiency. Source: Cooper et al. (2002a, b)

This can be illustrated by Fig. 9.1 where point C is technically inefficient but does not display the congestion that is associated with the negative sloped segment that connects point C to point D. For further details on this figure, see Chap. 7 in this handbook. The point C is evidently very sensitive to changes in its output (but not its input) data. In the terminology of Charnes et al. (1996), it has a "zero radius of stability." If its output is raised in any positive amount, C becomes efficient. Alternatively, if C is lowered it becomes an example of congestion. All these arguments are from a sensitivity analysis point of view. If we consider changes in point C to involve only random variations, these characterizations will change. They will not depend on the change in output level, but will depend rather on the specified probability level α. When $\alpha = 0.5$, random variations in the coordinates of point C do not have any impact on its efficiency, inefficiency, or congestion characterizations. Hence, it is satisfactory to employ the deterministic model (9.6) since, with this choice of α, the user is indifferent to the possible presence of inefficiency (or congestion) stochastically. This is different from the sensitivity analysis results. When α is taken between 0 and 0.5, point C will be efficient in the stochastic sense irrespective of the random variations (see Theorem 9.5(b), below). This is again different from the result of the sensitivity analysis discussed above. Finally, if α is assigned a value between 0.5 and 1, point C will be inefficient in the stochastic sense – no matter what the direction of random variations (see Theorem 9.6(b), below). Thus, in all cases the choice of α plays the critical role.

Theorem 9.4 (Cooper et al. 2002a, b). For $\alpha = 0.5$. The inefficiency vs. efficiency classification of DMU_o in input–output mean model (9.6) is the same as in stochastic model (9.46).

Theorem 9.5 (Cooper et al. 2002a, b). For $0 < \alpha < 0.5$.

(a) Suppose DMU_o is efficient with $DMU_o \in E \cup E'$ in input–output mean model (9.6), then $DMU_o \in E$ in stochastic model (9.46)
(b) Suppose $DMU_o \in F$ in input–output mean model (9.6), then $DMU_o \in E$ in stochastic model (9.46)

(c) Suppose $\mathrm{DMU_o} \in N$ in input–output mean model (9.6), then $\mathrm{DMU_o} \in N$ in stochastic model (9.46) if $\sigma_{i0}^{I} < \beta_i^{-*}/(-\Phi^{-1}(\alpha))$ and $\sigma_{r0}^{o} < \beta_r^{+*}/(-\Phi^{-1}(\alpha))$, where, for $\alpha < 0.5$ we have $\Phi^{-1}(\alpha) < 0$. Here $\sum_{r=1}^{s} \beta_r^{+*} + \sum_{i=1}^{m} \beta_i^{-*}$ is the optimal value of

$$
\begin{aligned}
&\text{Max} \sum_{r=1}^{s} \beta_r^{+} + \sum_{i=1}^{m} \beta_i^{-} \\
&\text{s.t.} \\
&\sum_{j=1}^{n} y_{rj}\lambda_j - \beta_r^{+} \geq y_{r0}, \qquad r = 1, \ldots, s, \\
&\sum_{j=1}^{n} x_{ij}\lambda_j + \beta_i^{-} \leq x_{i0}, \qquad i = 1, \ldots, m, \\
&\sum_{j=1}^{n} \lambda_j = 1, \\
&\beta_r^{+} \geq 0, \beta_i^{-} \geq 0, \lambda_j \geq 0, \quad r = 1, \ldots, s; \\
&\qquad\qquad\qquad\qquad\quad i = 1, \ldots, m; \quad j = 1, \ldots, n.
\end{aligned} \tag{9.49}
$$

Theorem 9.6 (Cooper et al. 2002a, b). For $1 > \alpha > 0.5$.

(a) Suppose $\mathrm{DMU_o} \in E$ in input–output mean model (9.6), then $\mathrm{DMU_o} \in E$ in stochastic model (9.46) if

$$
\sum_{r=1}^{s} \sigma_{r0}^{o} + \sum_{i=1}^{m} \sigma_{i0}^{I} < \left(\sum_{r=1}^{s} \theta_r^{+*} + \sum_{i=1}^{m} \theta_i^{-*} \right) / \Phi^{-1}(\alpha),
$$

where $\sum_{r=1}^{s} \theta_r^{+*} + \sum_{i=1}^{m} \theta_i^{-*}$ is the optimal value of

$$
\begin{aligned}
&\text{Min} \sum_{r=1}^{s} \theta_r^{+} + \sum_{i=1}^{m} \theta_i^{-} \\
&\text{s.t.} \\
&\sum_{j\neq 0}^{n} y_{rj}\lambda_j \geq y_{r0} - \theta_r^{+}, \quad r = 1, \ldots, s, \\
&\sum_{j\neq 0}^{n} x_{ij}\lambda_j \leq x_{i0} + \theta_i^{-}, \quad i = 1, \ldots, m, \\
&\sum_{j\neq 0}^{n} \lambda_j = 1, \\
&\theta_r^{+} \geq 0, \theta_i^{-} \geq 0, \lambda_j \geq 0 (j \neq 0), \quad r = 1, \ldots, s, \quad j = 1, \ldots, n, \\
&i = 1, \ldots, m.
\end{aligned} \tag{9.50}
$$

(b) Suppose $\mathrm{DMU_o} \in E' \cup F \cup N$ in input–output mean model (9.6), then $\mathrm{DMU_o} \in N$ in stochastic model (9.46).

9.5 Satisficing DEA Models

In the previous section, we have discussed joint chance-constrained DEA formulations and "E-model" chance-constrained DEA forms. In this section, we would like to discuss another type of chance-constrained DEA model. Referred to as "P-models" in the chance-constrained programming literature, we can also refer to them as "satisficing" DEA models, as drawn from Cooper et al. (1996).

We start by introducing the following version of a P-Model which we use to adapt the usual definitions of "DEA efficiency" to a chance-constrained programming context,

$$\begin{aligned}
&\text{Max } P\left\{\frac{\sum_{r=1}^{s} u_r \tilde{y}_{ro}}{\sum_{i=1}^{m} v_i \tilde{x}_{io}} \geq 1\right\} \\
&\text{s.t.} \\
&\{\text{P}\}\left\{\frac{\sum_{r=1}^{s} u_r \tilde{y}_{rj}}{\sum_{i=1}^{m} v_i \tilde{x}_{ij}} \leq 1\right\} \geq 1-\alpha_j, \quad j=1,\ldots,n, \\
&u_r, v_i \geq 0, \quad \forall r, i.
\end{aligned} \tag{9.51}$$

Here, P means "Probability" and the symbol ~ is used to identify the inputs and outputs as random variables with a known joint probability distribution. The u_r and $v_i \geq 0$ are the virtual multipliers (=weights) to be determined by solving the above problem. This model evidently builds upon the CCR model of DEA, as derived in Charnes et al. (1978), with the ratio in the objective representing output and input values for DMU_o, the DMU to be evaluated, which is also included in the $j = 1, \ldots, n$ DMUs with output-to-input ratios represented as chance constraints.

Evidently, the constraints in (9.51) are satisfied by choosing $u_r = 0$, and $v_i > 0$ for all r and i. Hence, for continuous distributions like the ones considered in this chapter, it is not vacuous to write,

$$P\left\{\frac{\sum_{r=1}^{s} u_r^* \tilde{y}_{ro}}{\sum_{i=1}^{m} v_i^* \tilde{x}_{io}} \leq 1\right\} + P\left\{\frac{\sum_{r=1}^{s} u_r^* \tilde{y}_{ro}}{\sum_{i=1}^{m} v_i^* \tilde{x}_{io}} \geq 1\right\} = 1$$

or

$$P\left\{\frac{\sum_{r=1}^{s} u_r^* \tilde{y}_{ro}}{\sum_{i=1}^{m} v_i^* \tilde{x}_{io}} \leq 1\right\} = 1-\alpha^* \geq 1-\alpha_0.$$

Here, * refers to an optimal value so α^* is the probability of achieving a value of at least unity with this choice of weights and $1-\alpha^*$ is, therefore, the probability of failing to achieve this value.

To see how these formulations may be used, we note that we must have $\alpha_0 \geq \alpha^*$ since $1-\alpha^*$ is prescribed in the constraint for $j=0$ as the chance allowed for characterizing the $\tilde{y}_{r0}$, $\tilde{x}_{i0}$ values as inefficient. More formally, we introduce the following stochasticized definition of efficiency.

Definition 9.11. $\mathrm{DMU_o}$ is "stochastic efficient" if and only if $\alpha^* = \alpha_0$.

This opens a variety of new directions for research and potential uses of DEA. Before indicating some of these possibilities, however, we replace (9.51) with the following:

$$\text{Max } P\left\{\frac{\sum_{r=1}^{s} u_r \tilde{y}_{ro}}{\sum_{i=1}^{m} v_i \tilde{x}_{io}} \geq 1\right\}$$

s.t.

$$P\left\{\frac{\sum_{r=1}^{s} u_r \tilde{y}_{rj}}{\sum_{i=1}^{m} v_i \tilde{x}_{ij}} \leq 1\right\} + P\left\{\frac{\sum_{r=1}^{s} u_r \tilde{y}_{ro}}{\sum_{i=1}^{m} v_i \tilde{x}_{io}} \geq 1\right\} \geq 1, \quad j = 1, \ldots, n,$$

$$u_r, v_i \geq 0, \quad \forall r, i. \tag{9.52}$$

This simpler model makes it easier to see what is involved in uses of these CCP/DEA formulations. It also enables us to examine potential uses in a simplified manner.

First, as is customary in CCP, it is assumed that the behavior of the random variables is governed by a known multivariate distribution. Hence, we can examine the value of α* even before the data are generated. If this value is too small then one can signal central management, say, that the situation for $\mathrm{DMU_o}$ needs to be examined in advance because there is a probability of at least $1-\alpha^* \geq 1-\alpha_0$ that it will not perform efficiently.

Some additional uses of these concepts can be brought into view from the original work in CCP. For instance, the article by Charnes et al. (1958) introduced CCP was concerned with policies and programs involved in scheduling heating oil production for EXXON (then known as Standard Oil of New Jersey). This led to the formation of a risk evaluation committee (the first in the Company's history) to determine suitable choices of α^*. It was decided that a "policy" to supply all customers on demand would require $\alpha^* > 1/2$ since the alternate choice of $\alpha^* \leq 1/2$ was likely to be interpreted by customers, and others, to mean that the company was either indifferent or unlikely to be willing to supply all customers on demand. This characterization and usage of the term "policy" was important because the company was especially concerned with heating oil as a "commodity charged with a public interest" since failure to supply it to customers on demand (in cold weather) could have severe consequences. See the discussion of this "policy" in Charnes et al. (1958).

If we define a "rule" as "a chance constraint which is to hold with probability one," then we can regard a "policy" as "a chance constraint which is to hold with probability $0.5<\alpha^*<1$." Implementation of a "policy" allows for deviations which can require managerial attention whereas a "rule" may be administered in clerical fashion since no exceptions are to be permitted. Notice, too, that a policy may be identified and evaluated by reference to ex post data, as in an accounting or performance audit, in order to see whether the corresponding actions had been taken sufficiently frequently, or whether some "policy" other than the intended one had prevailed, see Cooper and Ijiri (1983).

We can now bring the above discussion into focus for its possible use in efficiency evaluations because the constraint for $j = 0$ in (9.52) contains complementary possibilities. Hence, accepting a value of $\alpha^*>1/2$ means acceptance of a policy that favors efficient performance whereas a value of $\alpha^* \leq 1/2$ means that indifferent or inefficient performance is favored. This does not end the matter. The already calculated u_r^*, v_i^* remain available for use and may also be applied to the data that materialize after operations are undertaken by DMU_o. Applying the previously determined weights to the thus generated data allows us to calculate the probability that the values realized by DMU_o will occur. Using these weights, we may then determine whether the observed inputs and outputs yield a ratio that is within the allowable range of probabilities or whether a shift in the initially assumed multivariate distribution has occurred.

Further pursuit of this topic would lead into discussions of the higher order decision rules in CCP and/or the use of Bayesian procedures to modify the initially assumed probability distributions. See, e.g., R. Jaganathan (1985). We do not follow this route but prefer, instead, to move toward extensions of (9.51) that will enable us to make contact with the "satisficing" concepts of H. A. Simon (1957).

The following model represents an evident generalization of (9.51):

$$\text{Max } P\left\{\frac{\sum_{r=1}^{s} u_r \tilde{y}_{ro}}{\sum_{i=1}^{m} v_i \tilde{x}_{io}} \geq \beta_o\right\}$$

s.t.

$$P\left\{\frac{\sum_{r=1}^{s} u_r \tilde{y}_{rj}}{\sum_{i=1}^{m} v_i \tilde{x}_{ij}} \leq \beta_j\right\} \geq 1-\alpha_j, \quad j = 1, \ldots, n,$$

$$P\left\{\frac{\sum_{r=1}^{s} u_r \tilde{y}_{rj}}{\sum_{i=1}^{m} v_i \tilde{x}_{ij}} \geq \beta_j\right\} \geq 1-\alpha_j, \quad j = n+1, \ldots, n+k,$$

$$u_r, v_i \geq 0, \quad \forall r, i. \tag{9.53}$$

Here, we interpret the β_0 in the objective as an "aspiration level" either imposed by an outside authority, as in the budgeting model of A. Stedry (1960), or adopted by an individual for some activity as in the satisficing concept of H. A. Simon (1957). The α_j $(0 \leq \alpha_j \leq 1)$ in the constraints are predetermined scalars which represent an allowable chance (risk) of violating the constraints with which they are associated. The u_r, $v_i \geq 0$ are the virtual multipliers (=weights) to be determined by solving the above problem. This model evidently builds upon the "ratio form" for the CCR model of DEA with the ratio in the objective representing output and input values for DMU_o, the DMU to be evaluated. This ratio form is also included as one of the $j = 1, \ldots, n$ DMUs with output-to-input ratios represented as chance constraints. We may then think of the first $j = 1, \ldots, n$ constraints as representing various conditions such as physical possibilities or the endurance limits of this individual. The added constraints $j = n + 1, \ldots, n + k$ may represent further refinements of the aspiration levels. This could even include a constraint $\beta_j = \beta_0$ with a prescribed level of probability for achieving this aspired level that might exceed the maximum possible value. The problem would then have no solution. In such cases, according to Simon (1957), an individual must either quit or else he must revise his aspiration level or the risk of not achieving this level (or both). Thus, probabilistic (chance-constrained programming) formulations allow for types of behavior which are not present in the customary deterministic models of satisficing.

Uses of these ideas in actual applications are yet to be made. However, we think that potential uses include possibilities for using DEA to extend the kinds of behavior that are represented in the approaches used in both economics and psychology. For instance, it is now common to contrast "satisficing" and "optimizing behavior" as though the two are mutually exclusive. Reformulation and use of DEA along lines like we have just described, however, may enable us to discover situations in which both types of behavior might be present. Indeed it is possible that behaviors which are now characterized as inefficient (with deterministic formulations) might better be interpreted as examples of satisficing behavior with associated probabilities of occurrence. This kind of characterization may, in turn, lead to further distinctions in which satisficing gives way to inefficiencies when probabilities are too low even for satisficing behavior and this, we think, provides access to sharper and better possibilities than those offered in the economics literature by G. J. Stigler's (1976) critique of H. Leibenstein's (1976) concept of "X-Efficiency."

The preceding interpretations were pointed toward individual behaviors that accord with the satisficing characterizations provided in H. A. Simon (1957). Turning now to managerial uses, we can simplify matters by eliminating the last k constraints in (9.53) from consideration. One of the DMUs of interest is then singled out for evaluating the probability that its performance will exceed the $\beta_j = \beta_0$ value assigned to (or assumed for) this entity in the constraints. We can then interpret our results as being applicable in either an ex ante or ex post manner according to whether our interest is in planning or control. In a planning mode, for instance, we can determine a maximum probability of inefficient or satisficing

(tolerably inefficient) behavior that we may want to anticipate or forestall when $\alpha^* < \alpha_0$ occurs. For control purpose, we may similarly want to determine whether the observed behavior, as recorded, is too far out for us to regard it as having conformed to what should occur.

In a similar manner to the analysis of model (9.51), it is obvious that the first $j = 1, \ldots, n$ constraints in (9.53) are satisfied by choosing $u_r = 0$, and $v_i > 0$ for all r and i. For an optimal solution (u^*, v^*), we must have

$$P\left\{\frac{\sum_{r=1}^{s} u_r^* \tilde{y}_{ro}}{\sum_{i=1}^{m} v_i^* \tilde{x}_{io}} \leq \beta_0\right\} = 1 - \alpha* \geq 1 - a_0.$$

Therefore, we introduce the following stochasticized definitions of efficiency.

Definition 9.12 (Stochastic Efficiency). If $\beta_{j_0} = \beta_0 = 1$, $\mathrm{DMU_o}$ is "stochastically efficient" if and only if $\alpha^* = \alpha_0$.

Definition 9.13 (Satisficing Efficiency). If $\beta_{j_0} = \beta_0 < 1$, $\mathrm{DMU_o}$ is "satisficing efficient" if and only if $\alpha^* = \alpha_0$.

Following Land et al. (1992, 1993, 1994), we assume that input values are deterministic so that only the outputs are to be represented as random variables with a multivariate normal distribution and known parameters. Again like Land, Lovell, and Thore, we restrict attention to the class of zero-order decision rules.

The class of zero-order decision rules for use in chance-constrained programming can be most easily explained by turning to the example of scheduling heating oil production at EXXON where the objective was to secure a best schedule for this seasonal (weather dependent) product as required to anticipate the probabilistic demands. Decisions rules were developed that allowed for changing production schedules, in conditional stochastic fashion, as sales materialized. A zero-order rule, however, would have set the schedules for the entire season and use of this rule means that the vectors u and v of multipliers in (9.53) are to be treated as deterministic variables.

Choices of multivariate normal distributions and zero-order decision rules are less restrictive than might at first appear to be the case. Transformations are available for bringing other types of distributions into approximately normal form – as is done in Charnes et al. (1968), for instance, when it was found necessary to treat the case of highly skewed (log-normal) distributions which were encountered when developing new product marketing strategies. We can also adapt our use of zero-order decision rules by interpreting them as a series of one-period-at-a-time applications with appropriate models, to allow for changing realizations and probabilities, and regard these (in many situations) as approximations to the more complex solution procedures involved in developing higher order "conditional" decision rules to deal with full-scale treatment of the dynamics. Proceeding in this one-period-at-a-time manner also allows us to bypass additional problems such as

the sample size considerations which are encountered in dealing with multiple observations.

Now we assume $\alpha_j < 0.5$ (we will relax this later). Utilizing techniques in chance-constrained programming theory, a deterministic equivalent of (9.53) is then as follows:

$$\begin{aligned}
&\text{Max } \gamma \\
&\text{s.t.} \\
&u^T y_0 - \beta_0 v^T x_0 \geq \Phi^{-1}(\gamma)\sqrt{u^T \textstyle\sum_0 u} \\
&u^T y_j - \beta_j v^T x_j - \Phi^{-1}(\alpha_j)\eta_j \leq 0, \quad j = 1, \ldots, n, \\
&\eta_j^2 - u^T \textstyle\sum_j u \geq 0, \quad j = 1, \ldots, n, \\
&u \geq 0, v \geq 0, \eta \geq 0,
\end{aligned} \tag{9.54}$$

where $\sum_j = (\text{Cov}(\tilde{y}_{ij}, \tilde{y}_{kj}))$.

This is a nonlinear and nonconvex programming problem. However, let us consider the following quadratic programming problem:

$$\begin{aligned}
&\text{Max } \delta \\
&\text{s.t.} \\
&\mu^T y_0 - \beta_0 v^T x_0 \geq \delta, \\
&\mu^T \textstyle\sum_0 \mu \geq 1, \\
&\mu^T y_j - \beta_j v^T x_j - \Phi^{-1}(\alpha_j)\zeta_j \leq 0, \quad j = 1, \ldots, n, \\
&\zeta_j^2 - \mu^T \textstyle\sum_j \mu \geq 0, \quad j = 1, \ldots, n, \\
&\mu \geq 0, v \geq 0, \zeta \geq 0
\end{aligned} \tag{9.55}$$

It is easy to show that if δ^* is the optimal value of (9.55) and γ^* is the optimal value of (9.54), then we have $\Phi(\delta^*) = \gamma^*$. Therefore, we have the following result.

Theorem 9.7 (Cooper et al. 1996). If $\beta_{j_0} = \beta_0 = 1$ ($\beta_{j_0} = \beta_0 < 1$), DMU_o is stochastically (satisficing) efficient if and only if $\Phi(\delta^*) = \alpha_0$.

Now let us discuss the case of $\Phi(\delta^*) < \alpha_0$. In this case, the risk of failing to satisfy the constraints for DMU_{j_o} falls below the level which was specified as satisfactory. To state this in a more positive manner, let us consider the fact that

$$P\left\{\frac{\sum_{r=1}^{s} u_r^* \tilde{y}_{ro}}{\sum_{i=1}^{m} v_i^* x_{io}} \leq \beta_0\right\} + P\left\{\frac{\sum_{r=1}^{s} u_r^* \tilde{y}_{ro}}{\sum_{i=1}^{m} v_i^* x_{io}} \geq \beta_0\right\} = 1.$$

Therefore,

$$P\left\{\frac{\sum_{r=1}^{s} u_r^* \tilde{y}_{ro}}{\sum_{i=1}^{m} v_i^* x_{io}} \le \beta_0\right\} = 1 - \Phi(\delta^*) > 1 - \alpha_0.$$

This leads to the following corollary to the above theorem.

Corollary 9.1. If $\beta_{j_0} = \beta_0 = 1$ ($\beta_{j_0} = \beta_0 < 1$), DMU_o is stochastically (satisficing) inefficient if and only if $\Phi(\delta^*) < \alpha_0$.

Note: (1) if $\alpha_j = 0.5$ for DMU_j, the constraint $\zeta_j^2 - \mu^T \sum_j \mu \ge 0$ should be deleted from model (9.55); (2) if $\alpha_j > 0.5$ for DMU_j, the constraint $\zeta_j^2 - \mu^T \sum_j \mu \ge 0$ should be changed to be $\zeta_j^2 - \mu^T \sum_j \mu \le 0$.

9.6 Concluding Remarks

DEA with stochastic variations has recently received significant attention. Banker (1993), for example, incorporated statistical elements into DEA and developed a nonparametric approach with maximum likelihood methods to effect inferences in the presence of statistical noise. See the discussion in Chap. 11 of this handbook and an alternative which is based on "bootstrapping" in Chap. 10. Land et al. (1992, 1993, 1994) utilized the chance-constrained programming constraints (Charnes and Cooper 1959; Charnes et al. 1958) which they adapted to DEA. Olesen and Petersen (1995) developed a chance-constrained DEA model which uses piecewise linear envelopments of confidence regions for use with stochastic multiple inputs and multiple outputs. Cooper et al. (1996) incorporated Simon's (1957) "Satisficing Concepts" into DEA models with chance constraints in order (1) to effect contact with theories of behavior in social psychology as well as (2) to extend the potential uses of DEA models to cases where 100% efficiency can be replaced by aspired levels of performance. Cooper et al. (1998) further developed a "joint chance-constrained" DEA model to naturally generalize "Pareto-Koopmans Efficiency" to stochastic situations. Huang and Li (1996) utilized this joint chance-constrained concept to discuss general dominance structures in stochastic situations. Recently, Cooper et al. (2002a, b, 2003) have introduced chance-constrained models to deal with technical inefficiencies and congestion in stochastic situations. Additional stochastic DEA approaches and applications of chance-constrained DEA models can be found in, but are not limited to, Sengupta (1982, 1987, 1988, 1989, 1990), Olesen (2006), Post (2001), Sueyoshi (2000), Talluri et al. (2006), Wu and Olson (2008).

In the DEA literature dealing with chance-constrained programming, attention has been restricted to the class of "zero-order decision rules." This corresponds to a "here and now" approach to decision making in contrast to the "wait and see" approach that is more appropriate to dynamic settings in which it may be better to delay some parts of a decision until more information is available. To go further in this direction leads to

the difficult problem of choosing a problem that has only been addressed in any detail for the class of linear (first order) decision rules even in the general literature on chance-constrained programming and even this treatment was conducted under restrictive assumptions on the nature of random variables and their statistical distributions (Cooper et al. 2000). Finally, a major assumption on probability distributions is that they are known. Hence, there is a real need and interest in relaxing this assumption in future research (see Jagannathan 1985; Cooper et al. 2002a, b).

References

Aigner D, Lovell CAK, Schmidt P. Formulation and estimation of stochastic frontier production function models. J Econ. 1977;6:21–37.

Banker RD. Maximum likelihood, consistency and DEA: statistical foundations. Manag Sci. 1993;39:1265–73.

Banker RD, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manag Sci. 1984;30:1078–92.

Bowlin WF, Brennan J, Cooper WW, Sueyoshi T. DEA models for evaluating efficiency dominance, *Research Report*. Austin, TX: The University of Texas, Center for Cybernetic Studies; 1984.

Charnes A, Clarke RL, Cooper WW. An approach to testing for organizational slack via banker's game theoretic DEA formulations. Res Govern Non Profit Account. 1989;5:211–29.

Charnes A, Cooper WW. Chance constrained programming. Manag Sci. 1959;5:73–9.

Charnes A, Cooper WW, DeVoe JK, Learner DB. Demon, mark II: an extremal equation approach to new product marketing. Manag Sci. 1968;14:513–24.

Charnes A, Cooper WW, Golany B, Seiford L, Stutz J. Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. J Econ. 1985;30:91–107.

Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res. 1978;2:429–42.

Charnes A, Cooper WW, Rhodes E. Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through in US public school education. Manag Sci. 1981;27:668–97.

Charnes A, Cooper WW, Symonds GH. Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil. Manag Sci. 1958;4:235–63.

Charnes A, Cooper WW, Thrall RM. Identifying and classifying efficiencies and inefficiencies in data envelopment analysis. Oper Res Lett. 1986;5:105–16.

Charnes A, Cooper WW, Thrall RM. A structure for classifying efficiencies and inefficiencies in data envelopment analysis. J Product Anal. 1991;2:197–237.

Charnes AS, Haag PJ, Semple J. Sensitivity of efficiency classifications in the additive model of data envelopment analysis. Int J Syst Sci. 1992;23:789–98.

Charnes A, Neralic L. Sensitivity analysis of the additive model in data envelopment analysis. Eur J Oper Res. 1990;48:332–41.

Charnes A, Rousseau J, Semple J. Sensitivity and stability of efficiency classifications in data envelopment analysis. J Product Anal. 1996;7:5–18.

Clarke RL. Effects of repeated applications of data envelopment analysis on efficiency of air force vehicle maintenance units in the tactical air command. Ph.D. Thesis. The University of Texas at Austin, Graduate School of Business; 1989. Also available from University Microfilms, Inc., in Ann Arbor, Michigan.

Cooper WW, Deng H, Huang ZM, Li SX. Chance constrained programming approaches to technical efficiencies and inefficiencies in stochastic data envelopment analysis. J Oper Res Soc. 2002a;53:1347–56.
Cooper WW, Deng H, Huang ZM, Li SX. Chance constrained programming approaches to congestion in stochastic data envelopment analysis. Eur J Oper Res. 2003;155:231–8. 2002.
Cooper WW, Huang ZM, Lelas V, Li SX, Olesen OB. Chance constrained programming formulations for stochastic characterizations of efficiency and dominance in DEA. J Product Anal. 1998;9:53–79.
Cooper WW, Ijiri Y, editors. Kohler's Dictionary for Accountants. 6th ed. Englewood Cliffs, NJ: Prentice Hall, Inc; 1983.
Cooper WW, Lelas V, Sullivan DW. Chance-constrained programming and skewed distributions of matrix coefficients and applications to environmental regulatory activities. In: Cawrence KD, Kleinbore RK, editors. Science: in productivity, finance and management. Oxford: Elsevier Publishers; 2002b.
Cooper WW, Li S, Seiford LM, Tone K, Thrall RM, Zhu J. Sensitivity and stability analysis in DEA: some recent developments. J Product Anal. 2001;15:217–46.
Cooper WW, Huang ZM, Li SX. Satisficing DEA models under chance constraints. Ann Oper Res. 1996;66:279–95.
Cooper WW, Seiford LM, Tone K. Data envelopment analysis: a comprehensive text with models, applications, reference and DEA-solver software. Boston, MA: Kluwer; 2000.
Huang ZM, Li SX. Dominance stochastic models in data envelopment analysis. Eur J Oper Res. 1996;95:370–403.
Jagannathan R. Use of sample information in stochastic recourse and chance constrained programming. Manag Sci. 1985;31:96–108.
Land KC, Lovell CAK, Thore S. Productivity and efficiency under capitalism and state socialism: the chance-constrained programming approach. Proceedings of the 47th Congress of the International Institute of Public Finance, Pestieau P. (ed.), 1992. 109–21.
Land KC, Lovell CAK, Thore S. Chance-constrained data envelopment analysis. Manag Decis Econ. 1993;14:541–54.
Land KC, Lovell CAK, Thore S. Productivity and efficiency under capitalism and state socialism: an empirical inquiry using chance-constrained data envelopment analysis. Technol Forecast Soc Chang. 1994;46:139–52.
Leibenstein H. Beyond economic man. Cambridge: Harvard University Press; 1976.
Olesen OB, Petersen NC. Chance constrained efficiency evaluation. Manag Sci. 1995;41:442–57.
Olesen OB. Comparing and combining two approaches for chance constrained DEA. J Product Anal. 2006;26:103–19.
Post T. Performance evaluations in stochastic environments using mean-variance data envelopment analysis. Oper Res. 2001;49:281–92.
Seiford LM, Zhu J. Stability regions for maintaining efficiency in data envelopment analysis. Eur J Oper Res. 1998;108:127–39.
Sengupta JK. Efficiency measurement in stochastic input–output systems. Int J Syst Sci. 1982;13:273–87.
Sengupta JK. Data envelopment analysis for efficiency measurement in the stochastic case. Comput Oper Res. 1987;14:117–29.
Sengupta JK. Robust efficiency measures in a stochastic efficiency. Int J Syst Sci. 1988;19:779–91.
Sengupta JK. Data envelopment with maximum correlation. Int J Syst Sci. 1989;20:2085–93.
Sengupta JK. Transformations in stochastic DEA models. J Econ. 1990;46:109–24.
Simon HA. Models of man. New York: Wiley; 1957.
Sinha KK. Moving frontier analysis: an application of data envelopment analysis for competitive analysis of a high technology manufacturing plant. Ann Oper Res. 1996;66:197–218.
Stedry AC. Budget control and cost behavior. Englewood Cliffs, NJ: Prentice-Hall; 1960.
Stigler GJ. The existence of X-efficiency. Am Econ Rev. 1976;66:213–6.

Sueyoshi T. Stochastic DEA for restructure strategy: an application to a Japanese Petroleum Company. Omega. 2000;28:385–98.
Talluri S, Narasimhan R, Nair A. Vendor performance with supply risk: a chance-constrained DEA approach. Int J Prod Econ. 2006;100:212–22.
Wu D, Olson DL. A comparison of stochastic dominance and stochastic DEA for vendor evaluation. Int J Prod Res. 2008;46:2313–27.

Chapter 10
Performance of the Bootstrap for DEA Estimators and Iterating the Principle

Léopold Simar and Paul W. Wilson

Abstract This chapter further examines the bootstrap method proposed by Simar and Wilson (Manag Sci 44(11):49–61, 1998) for DEA efficiency estimators. Some simplifications as well as Monte Carlo evidence on the coverage probabilities of confidence intervals estimated by the method are offered. In addition, we present similar evidence for confidence intervals estimated with the so-called naive bootstrap to illustrate the fact that the naive bootstrap is inconsistent in the DEA setting. Finally, we propose an iterated version of the bootstrap which may be used to improve bootstrap estimates of confidence intervals.

Keywords Data envelopment analysis • Bootstrap • Distance function • Efficiency • Frontier models

10.1 Introduction

Nonparametric efficiency estimators such as data envelopment analysis (DEA) typically rely on linear programming (LP) techniques for computation of estimates, and are often characterized as *deterministic* (as opposed to *econometric* or *statistical*), as if to suggest that the methods lack any statistical underpinnings. Applied studies that have used these methods have typically presented point estimates of inefficiency, with no measure or even discussion of uncertainty surrounding these estimates. Indeed, many papers contain statements where efficiency is described as being *computed* or *calculated* as opposed to being *estimated*, and results are frequently referred to as *efficiencies* rather than *efficiency estimates*.

The choice of terminology in describing the nonparametric efficiency approaches and their results is perhaps understandable given the (until very

P.W. Wilson (✉)
Department of Economics, Clemson University, Clemson, SC, USA
e-mail: pww@clemson.edu

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_10,

recently) lack of a "tool box" with aids for diagnostics, inference, etc., such as the one available to researchers using the parametric approaches. But the terminology is also unfortunate, because it has served to cloud important issues in efficiency estimation.

Today, researchers have access to a growing set of tools for statistical inference within nonparametric efficiency estimation; these tools are based on bootstrap methods. This chapter summarizes some of those tools, and shows how the bootstrap principle may be iterated to improve confidence interval estimates. Section 10.2 briefly reviews the microeconomic theory of the firm, and establishes our notation and an economic model. Section 10.3 discusses the estimation of efficiency, while Sect. 10.4 establishes a statistical model by augmenting the economics assumptions in Sect. 10.2 with some assumptions on the data-generating process (DGP). Section 10.5 briefly reviews existing asymptotic results, which have implications for efficiency of estimation as well as for the bootstrap. The DEA bootstrap is discussed generally in Sect. 10.6, and more specifically in Sect. 10.7, where we give details on its implementation. Section 10.8 details our Monte Carlo experiments and their results. Section 10.9 offers an iterated version of the bootstrap which can be used to improve estimates of confidence intervals. The final section offers some concluding remarks.

10.2 Efficiency and the Theory of the Firm

Standard microeconomics texts develop the theory of the firm by positing a production set which describes how a set of inputs may be somehow converted into outputs. To illustrate, let $x \in R_+^p$ denote a vector of p inputs and $y \in R_+^q$ denote a vector of q outputs. Then the production set may be defined as

$$P \equiv \{(x, y) | x \text{ can produce } y\}, \tag{10.1}$$

which is merely the set of feasible combinations of x and y. The production set P is sometimes described in terms of its sections

$$Y(x) \equiv \{y | (x, y)| \in P\}, \tag{10.2}$$

and

$$X(y) \equiv \{x | (x, y)| \in P\}, \tag{10.3}$$

which form the output feasibility and input requirement sets, respectively. Knowledge of either $Y(x)$ for all x or $X(y)$ for all y is equivalent to knowledge of P; P implies (and is implied by) both $Y(x)$ and $X(y)$. Thus, both $Y(x)$ and $X(y)$ inherit the properties of P. Various assumptions regarding P are possible; we adopt those of Shephard (1970) and Färe (1988).

Assumption A1. P is convex, $Y(x)$ is convex, bounded, and closed for all $x \in R_+^p$, and $X(y)$ is convex, bounded, and closed for all $y \in R_+^q$.

Assumption A2. $(0, y) \notin P$ if $y \geq 0, y \neq 0$, i.e., all production requires use of some inputs.

Assumption A2 merely says that there are no free lunches.[1]

Assumption A3. For $\tilde{x} \geq x, \tilde{y} \leq y,\ \tilde{x} \in R_+^p,\ \tilde{y} \in R_+^q$, if $(x, y) \in P$ then $(\tilde{x}, \tilde{y}) \in P$, i.e., both inputs and outputs are strongly disposable.

Assumption A3 is sometimes called free disposability. The presentation in this section does not depend on these particular assumptions, and the estimators discussed in Sect. 10.3 can be adapted to cases where alternative assumptions might be warranted. For example, if pollution is an inadvertent byproduct of the production process, then it might not be reasonable to assume that this particular output is strongly (freely) disposable.

The boundary of P is sometimes referred to as the *technology* or *production frontier*, and is given by the intersection of P and the closure of its complement, P^{∂}. Similarly, isoquants are defined by the boundary of $X(y)$,

$$X^{\partial}(y) = \{x | x \in X(y), \theta x \notin X(y)\, \forall\, 0 < \theta < 1\}, \tag{10.4}$$

while iso-output curves are defined by the boundary of $Y(x)$,

$$Y^{\partial}(x) = \{y | y \in Y(x), \lambda y \notin Y(x)\, \forall\, \lambda > 1\}. \tag{10.5}$$

Firms which are *technically inefficient* operate at points in the interior of P, while those that are technically *efficient* operate somewhere along the technology defined by P^{∂}. Various measures of technical efficiency are possible. The Shephard (1970) output distance function provides a normalized measure of Euclidean distance from a point $(x, y) \in P$ to P^{∂} in a direction orthogonal to x, and may be defined as

$$D(x, y|P) \equiv \inf\left\{\theta > 0 \middle| (x, \theta^{-1} y)\right| \in P\}. \tag{10.6}$$

Clearly, $D(x, y|P) \leq 1$ for all $(x, y) \in P$. If $D(x, y|P) = 1$ then $(x, y) \in P^{\partial}$; i.e., that the point (x, y) lies on the boundary of P, and the firm is technically efficient. One can similarly define the Shephard (1970) input distance function, which provides a normalized measure of Euclidean distance from a point $(x, y) \in P$ to P^{∂} in a direction orthogonal to y. Alternatively, one could employ measures of hyperbolic graph efficiency defined by Färe et al. (1985), directional distance functions as defined by Chambers et al. (1996), or perhaps other measures. From a purely technical viewpoint, any of these can be used to measure technical

[1] Throughout, inequalities involving vectors are defined on an element-by-element basis; e.g., for, $\tilde{x}, x \in R_+^p$ means that some, but perhaps not all or none, of the corresponding elements of $\tilde{x}$ and x may be equal, while some (but perhaps not all or none) of the elements of $\tilde{x}$ may be greater than corresponding elements of x.

efficiency; the only real difference is in the direction in which distance to the technology is measured. However, if one wishes to take account of behavioral implications, then this might influence the choice between the various possibilities. We present our methodology only in terms of the output orientation to conserve space, but our discussion can be adapted to the other cases by straightforward changes in our notation. Note the Shephard output distance function is the reciprocal of the Farrell (1957) output efficiency measure, while the Shephard input distance function is the reciprocal of the Farrell input efficiency measure.

In addition to technical efficiency, one may consider whether a particular firm is allocatively efficient. Assuming the firm is technically efficient, its input-allocative efficiency depends on whether it operates at the optimal location along one of the isoquants $X^{\partial}(y)$ as determined by prevailing input prices. Alternatively, output-allocative efficiency depends on whether the firm operates at the optimal location along one of the iso-output curves $Y^{\partial}(x)$ as determined by prevailing output prices. Färe et al. (1985) combine these notions of allocative efficiency with technical efficiency to obtain measures of "overall" efficiency. See Sect. 1.5 where these relations between technical and overall inefficiency are discussed. Here, we focus only on output technical efficiency, but it is straightforward to extend our discussion to the other notions of inefficiency.

Standard microeconomic theory suggests that with perfectly competitive input and output markets, firms which are either technically or allocatively inefficient will be driven from the market. However, in the real world, even where markets may be highly competitive, there is no reason to believe that this must happen instantaneously. Indeed, due to various frictions and imperfections in real markets for both inputs and outputs, this process might take many years, and firms that are initially inefficient in one or more respects may recover and begin to operate efficiently before they are driven from the market. Wheelock and Wilson (1995, 2000) provide support for this view through empirical evidence for banks operating in the USA.

10.3 Estimation

Unfortunately, none of the theoretical items defined in the previous section are observed, including the production set P and the output distance function $D(x, y|P)$. Consequently, these must be *estimated*.

In the typical situation, all that is observed are inputs and outputs for a set of n firms; together, these comprise the observed sample:

$$S_n = \{(x_i, y_i)\}_{i=1}^{n}. \tag{10.7}$$

These may be used to construct various estimators of P, which in turn may be used to construct estimators of $D(x, y|P)$. Charnes et al. (1978) used the conical hull

of the free disposal hull (FDH) of S_n to construct an estimator $\hat{V}$ of P in what has been called the "CCR model." This approach implicitly imposes an assumption of constant returns to scale on the production technology, P^{∂}. Banker et al. (1984) used the convex hull of the FDH of S_n as an estimator $\hat{P}$ of P. This approach, which has been termed the "BCC model," allows P^{∂} to exhibit increasing, constant, or decreasing returns at different locations along the frontier; hence $\hat{P}$ is said to allow variable returns to scale. Other estimators of P are possible, such as the FDH of S_n by itself (denoted $\tilde{P}$) as in Deprins et al. (1984), which in effect relaxes the assumption of convexity on P.

Terminology used in the DEA literature is sometimes rather confusing and misleading, and the preceding discussion offers an example. The terms "CCR model" and "BCC model" are misnomers; the conical and convex hulls of the FDH of S_n are merely *different estimators* of P – not different models. Similarly, the term "DEA model" appears far too frequently in the literature and serves only to obfuscate the fact that DEA is a class of estimators (characterized by, among other things, convexity assumptions). By contrast, a (statistical) model consists of a set of assumptions on (1) the underlying distribution from which the data are drawn and its support and (2) the process by which data are sampled (e.g., independently or otherwise) from this distribution (see Spanos 1986, for additional discussion). Together, these assumptions define a DGP. Given any one of these estimators of P, estimators of $D(x,y|P)$ can be constructed. For example, using $\hat{P}$, we may write

$$D(x,y|\hat{P}) \equiv \inf\{\theta > 0 | (x,\theta^{-1}y) \in \hat{P}\}, \tag{10.8}$$

where $\hat{P}$ has replaced P on the right-hand side of (10.6). Alternative estimators of $D(x,y|P)$ can be obtained by defining

$$D(x,y|\hat{V}) \equiv \inf\{\theta > 0 | (x,\theta^{-1}y) \in \hat{V}\}, \tag{10.9}$$

and

$$D(x,y|\tilde{P}) \equiv \inf\{\theta > 0 | (x,\theta^{-1}y) \in \tilde{P}\}. \tag{10.10}$$

The estimators in (10.8) and (10.9) can be used to construct tests of returns to scale; see Simar and Wilson (2002) and the discussion in Chap. 2.

It is similarly straightforward (with the benefit of hindsight and the pioneering work of Charnes et al. 1978) to compute estimates of $D(x,y|P)$ using (10.8)–(10.10) and the data in S_n. In particular, (10.8)–(10.9) may be rewritten as linear programs:

$$\left[D(x,y|\hat{P})\right]^{-1} = \max\{\varphi | \varphi y \leq Y\lambda, x \geq X\lambda, i\lambda = 1, \lambda \in R^n_+.\}, \tag{10.11}$$

and

$$\left[D(x,y|\hat{V})\right]^{-1} = \max\{\varphi | \varphi y \leq Y\lambda, x \geq X\lambda, \lambda \in R^n_+\}, \tag{10.12}$$

where $Y = [y_1, \ldots, y_n]$, $X = [x_1, \ldots, x_n]$, with each $x_i, y_i, \quad i = 1, \ldots, n$ denoting the $(p \times 1)$ and $(q \times 1)$ vectors of observed inputs and outputs (respectively), i is a $(1 \times n)$ vector of ones, and $\lambda = [\lambda_1, \ldots, \lambda_n]'$ is a $(n \times 1)$ vector of intensity variables. One can also rewrite (10.10) as a linear program, e.g.,

$$\left[D(x, y|\tilde{P})\right]^{-1} = \max\{\varphi | \varphi y \leq Y\lambda,\ x \geq X\lambda, \quad \lambda \in \{0, 1\}\}, \tag{10.13}$$

although linear programming is typically not used to compute estimates for the estimator based on the FDH.

Given our assumption that P is convex, by construction,

$$\tilde{P} \subseteq \hat{P} \subseteq \begin{cases} \hat{V} \\ P \end{cases}. \tag{10.14}$$

If the technology P^{∂} exhibits constant returns to scale everywhere, then it is also the case that $\hat{V} \subseteq P$, but otherwise, $\hat{V} \not\subset P$.

The differences between P and any of the estimators $\hat{P}$, $\hat{V}$, and $\tilde{P}$ are of utmost importance, for these differences determine the difference between $D(x, y|P)$ and any of the corresponding estimators $D(x, y|\hat{P})$, $D(x, y|\hat{V})$, and $D(x, y|\tilde{P})$. Note that P and especially $D(x, y|P)$ are the *true* quantities of interest. By contrast, $\hat{P}$, $\hat{V}$, and $\tilde{P}$ are *estimators* of P, while $D(x, y|\hat{P})$, $D(x, y|\hat{V})$, and $D(x, y|\tilde{P})$ are each *estimators* of $D(x, y|P)$. The true things that we are interested in, namely P and $D(x, y|P)$, are fixed, but unknown. Estimators, on the other hand, are necessarily *random variables*. When the data in S_n are used together with (10.11)–(10.13) to compute numerical values for the estimators $D(x, y|\hat{P})$, $D(x, y|\hat{V})$, and $D(x, y|\tilde{P})$, the resulting real numbers are merely specific realizations of different random variables; these numbers (which are typically those reported in efficiency studies) are *estimates*.

In our view, anything we might compute from data is an estimate. The formula used for such a computation defines an estimator. The DEA setting is not an exception to this principle.

The skeptical reader may well ask, "what if I have observations on *all* the firms in the industry I am examining?" This is a perfectly reasonable question, but it begs for additional questions. What happens if an entrepreneur establishes a new firm in this industry? How well might this new firm perform? Is there any reason why, at least in principle, it cannot perform better than the original firms in the industry? In other words, is it reasonable to assume that the new firm cannot operate somewhere above the convex hull of the existing firms represented by input/output vectors in R_+^{P+q}?

We can see no reason why the general answer to this last question must be anything other than a resounding "no!" To answer otherwise would be to deny the existence of technical progress, and by extension the historical experience of the civilized world over the last several hundred years. If one accepts this answer, it seems natural then to think in terms of a conceptual infinite population of potential firms, and to view one's sample as having resulted from a draw from an infinite population.

10.4 A Statistical Model

Continuing with the reasoning at the end of the previous section, once one accepts that the world is an uncertain place and all one can hope for are reliable estimators of unobservables such as $D(x, y|P)$, some additional questions are raised:

- What are the properties of the estimators defined in (10.8)–(10.10)? Are they at least consistent? If not, then even with an infinite amount of data, one might have little hope of obtaining a useful estimate.
- How much can be learned about the true value of interest? Is inference possible?
- Can hypotheses regarding P^{∂} be tested, even though P^{∂} is unobservable?
- Will inferences and hypothesis tests be reliable, meaningful?

Before these questions can be answered, a statistical model must be defined. A statistical model is merely a set of assumptions on the DGP. While many assumptions are possible, prudence dictates that the assumptions be chosen to provide enough structure so that desirable properties of estimators can be derived, but without imposing unnecessary restrictions. In addition to Assumptions A1–A3 adopted in Sect. 10.2, we now adopt assumptions based on those of Kneip et al. (1998).

Assumption A4. The sample observations in S_n are realizations of identically, independently distributed (iid) random variables with probability density function $f(x, y)$ with support over P.

Note that a point $(x, y) \in R_+^{p+q}$ represented by Cartesian coordinates can also be represented by cylindrical coordinates (x, ω, η) where (ω, η) are the polar coordinates of y in R_+^q. Thus, the modulus $\omega = \omega(y) \in R_+^1$ of y is given by the square root of the sum of the squared elements of y, and the jth element of the corresponding angle $\eta = \eta(y) \in [0, \pi/2]^{q-1}$ of y is given by arctan (y^{j+1}/y^1) for $y^1 \neq 0$ (where y^j represents the jth element of y); if $y^1 = 0$, then all elements of $\mu(y)$ equal $\pi/2$.

Writing $f(x, y)$ in terms of the cylindrical coordinates, we can decompose the density by writing

$$f(x, \omega, \eta) = f(\omega|x, \eta)\, f\,(\eta|x)\, f\,(x), \tag{10.15}$$

where all the conditional densities exist. In particular, $f(x)$ is defined on R_+^p, $f(\eta|x)$ is defined on $[0, \pi/2]^{q-1}$, and $f(\omega|x, \eta)$ is defined on R_+^1. Now consider a point $(x, y) \in P$, and the corresponding point $(x, y^{\partial}(x))$ which is the projection of (x, y) onto P^{∂} in the direction orthogonal to x, i.e., $y^{\partial}(x) = y/D(x, by|P)$. Then the moduli of these points are related to the output distance function via

$$0 \leq D(x, y|P) = \frac{\omega(y)}{\omega(y^{\partial}(x))} \leq 1. \tag{10.16}$$

Then the density $f(\omega|x,\eta)$ on $[0,\omega(y^{\partial}(x))]$ implies a density $f(D(x,y|P)|x,\eta)$ on the interval [0,1].

In order for our estimators of P and $D(x,y|P)$ to be consistent, the probability of observing firms in a neighborhood of P^{∂} must approach unity as the sample size increases.

Assumption A5. For all $x \geq 0$ and all $\eta \in [0,\pi/2]^{q-1}$, there exist constants $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that for all $\omega \in \left[\omega(y^{\partial}(x)), \omega(y^{\partial}(x)) + \varepsilon_2\right], f(\omega|x,\eta) \geq \varepsilon_1$.

In addition, an assumption about the smoothness of the frontier is needed[2]:

Assumption A6. For all (x,y) in the interior of P, $D(x,y|P)$ is differentiable in both its arguments.

Assumptions A1–A6 define the DGP F which yields the data in S_n. These assumptions are somewhat more detailed than the set of assumptions listed in Sect. 10.2, which were motivated by microeconomic theory.

10.5 Some Asymptotic Results

Recent papers have provided some asymptotic results for DEA/FDH estimators. Korostelev et al. (1995a, b) examined the convergence of estimators of P, and proved (for the special case where $p \geq 1$, $q = 1$) that both $\tilde{P}$ and $\hat{P}$ are consistent estimators of P in the sense that

$$\mathrm{d}\left(\tilde{P}, P\right) = O_p\left(n^{-\frac{1}{p+1}}\right), \tag{10.17}$$

and

$$\mathrm{d}\left(\hat{P}, P\right) = O_p\left(n^{-\frac{2}{p+2}}\right), \tag{10.18}$$

where $\mathrm{d}(\hat{P}, P)$ is the Lebesgue measure of the difference between two sets.[3]

The result for $\tilde{P}$ in (10.17) still holds if assumption A1 given above is dropped. However, the convexity assumption is necessary for (10.18). It would be seemingly straightforward to extend these results to obtain a convergence rate for $\hat{V}$ provided one adopted an additional assumption that P^{∂} displays constant returns to scale everywhere.

[2] Our characterization of the smoothness condition here is stronger than required; Kneip et al. (1998) require only Lipschitz continuity for $D(x,y|P)$, which is implied by the simpler, but stronger requirement presented here.

[3] Banker (1993) showed, for the case $q = 1, p \geq 1$, that $\hat{P}$ is a consistent estimator of P, but did not provide convergence rates.

These results also indicate that the *curse of dimensionality* which typically plagues nonparametric estimators is at work in the DEA/FDH setting; the convergence rates decrease as the number of inputs is increased. The practical implication of this is that researchers will need increasing amounts of data to get meaningful results as the number of inputs is increased. And, although these results were obtained with $q = 1$, presumably allowing the number of outputs to increase would have a similar effect on convergence rates and the resulting data requirements.

This intuition is confirmed by Kneip et al. (1998) and Park et al. (1999), who derive convergence rates for $D(x, y|\hat{P})$ and $D(x, y|\tilde{P})$, respectively, for the general case where $p \geq 1$, $q \geq 1$:

$$D(x, y|\hat{P}) - D(x, y|P) = O_p\left(n^{-\frac{2}{p+q+1}}\right) \tag{10.19}$$

and

$$D(x, y|\tilde{P}) - D(x, y|P) = O_p\left(n^{-\frac{1}{p+q}}\right). \tag{10.20}$$

In both cases, the convergence rates are affected by $p + q$. The convergence rate for the FDH estimator is slower than that for the convex hull estimator, and thus the convex hull estimator is preferred when P is convex. But if Assumption A1 does not hold, then there is no choice but to use the FDH estimator – the convex hull estimator as well as the convex cone estimator would be inconsistent in this case since $\tilde{P}$ and $\tilde{V}$ are convex for all $1 \leq n \leq \infty$, and so cannot converge to a nonconvex set as $n \to \infty$.

For the special case where $p = q = 1$, Gijbels et al. (1999) derived results which indicate that

$$D(x, y|\hat{P}) - D(x, y|P) \overset{\text{asy.}}{\sim} G(\cdot, \cdot), \tag{10.21}$$

where G is a known distribution function depending on unknown quantities related to the DGP F.[4]

For the general case where $p \geq 1$ and $q \geq 1$, Park et al. (1999) prove that

$$D(x, y|\tilde{P}) - D(x, y|P) \overset{\text{asy.}}{\sim} \text{Weibull}(\cdot, \cdot), \tag{10.22}$$

[4] In particular, the unknown quantities are determined by the curvature of P^{∂} and the value of $f(x, y)$ at the point where (x, y) is projected onto P^{∂} in the direction orthogonal to x. See Gijbels et al. (1999) for additional details.

where the unknown parameters of the Weibull distribution again must be estimated and depend on the characteristics of the DGP.

For the (variable returns to scale) DEA estimator, with $p \geq 1$ and $q \geq 1$, Kneip et al. (2008) derive the limiting distribution in the case of variable returns to scale with input or output orientation, while Wilson (2010) derived similar results for a hyperbolic version of the DEA estimator. In addition, Kneip et al. prove the consistency of two bootstrap methods, one smoothing not only the distribution of the data, but also the initial DEA frontier estimate. The other method involves subsampling, where bootstrap samples of size m are drawn, with $m < n$, $m \to \infty$, and $m/n \to 0$ as $n \to \infty$. The limiting distribution of the DEA efficiency estimator is quite complicated and involves unknown parameters; hence the bootstrap remains the only viable method for inference about efficiency levels. Park et al. (2010) have similarly derived the limiting distribution of the DEA efficiency estimator in the case of constant returns to scale.

These results are potentially very useful for applied researchers. In particular, (10.22) may be used to construct and then estimate confidence intervals for $D(x, y|P)$ when the FDH estimator is used. It is unfortunate that a similar result for the DEA case to date exists only for the special case where $p = q = 1$. In any case, these are asymptotic results, and the quality of confidence intervals, etc., estimated using these results in small samples remains an open question.

10.6 Bootstrapping in DEA/FDH Models

The bootstrap (Efron 1979, 1982) offers an alternative approach to classical inference and hypothesis testing. In the case of DEA estimators with multiple inputs or outputs, the bootstrap currently offers the *only* sensible approach to inference and hypothesis testing.

Bootstrapping is based on the analogy principle (e.g., see Manski 1988). The presentation here is based on our earlier work in Simar and Wilson (1998, 1999c). In the *true world*, we observe the data in S_n which are generated from

$$F = F(P, f(x, y)). \tag{10.23}$$

In the true world, F, P, and $D(x, y|P)$ are unobserved, and must be estimated using S_n.

Let $\hat{F}(S_n)$ be a consistent estimator of F:

$$\hat{F}(S_{n)} = F(\hat{P}, \hat{f}(x, y)). \tag{10.24}$$

One can easily simulate what happens in the true world by drawing a new dataset $S_n^* = \{x_i^*, y_i^*\}_{i=1}^n$ (analogous to (10.7)) from $\hat{F}(S_n)$, and then applying the original estimator to these new data. If the original estimator for a point $(x_{0,}\, y_0)$ (not

necessarily contained in S_n) was $D(x_0, y_0|\hat{P})$, then the estimator obtained from $\hat{F}(S_n)$ would be $D(x_0, y_0|\hat{P}^*)$, where $\hat{P}^*$ denotes the convex hull of the FDH of S_n^*, and is obtained by solving

$$\left[D(x_0, y_0|\hat{P}^*)\right]^{-1} = \max\{\varphi|\varphi y_0 \leq Y^*\lambda, x_0 \geq X^*\lambda, i\lambda = 1, \quad \lambda \in R_+^n\}, \quad (10.25)$$

where $Y^* = [y_1^*, \ldots, y_n^*]$, $X^* = [x_1^* \ldots x_n^*]$, with each, (x_i^*, y_i^*), $i = 1, \ldots, n$, denoting observations in the pseudodataset S_n^*. Repeating this process B times (where B is appropriately large) will result in a set of bootstrap values $\left\{D_b(x_0, y_0|\hat{P}^*)\right\}_{b=1}^B$. When the bootstrap is consistent, then

$$\begin{aligned}&\left(D(x_0, y_0|\hat{P}^*) - D(x_0, y_0|\hat{P})\right)|F(\hat{P}, \hat{f}(x, y)) \overset{\text{approx}}{\sim} \\ &\left(D(x_0, y_0|\hat{P}) - D(x_0, y_0|P)\right)|F(P, f(x, y)). \end{aligned} \quad (10.26)$$

Given the original estimate $D(x_0, y_0|\hat{P})$ and the set of bootstrap values

$$\left\{D_b(x_0, y_0|\hat{P}^*)\right\}_{b=1}^B,$$

the left-hand side of (10.26) is known with arbitrary precision (determined by the choice of number of bootstrap replications, B); the approximation improves as the sample size, n, increases.

Note that we could tell the story of the previous paragraph in terms of the convex cone estimator by merely changing $\hat{P}$ to $\hat{V}$. Equation (10.26) is the essence of the bootstrap. In principle, since $F(\hat{P}, \hat{f}(x, y))$ is known, it should be possible to determine the distribution on the left-hand side analytically. In practice, however, this is intractable in most problems, including the present one. Hence, Monte Carlo simulations are used to approximate this distribution. In our presentation, after the B bootstrap replications are performed, the set of bootstrap values $\left\{D_b(x_0, y_0|\hat{P}^*)\right\}_{b=1}^B$ gives an empirical approximation to this distribution.

Once the set of bootstrap values $\left\{D_b(x_0, y_0|\hat{P}^*)\right\}_{b=1}^B$ has been obtained, it is straightforward to estimate confidence intervals for the true distance function value $D(x_0, y_0|P)$. This is accomplished by noting that *if* we knew the true distribution of

$$\left(D(x_0, y_0|\hat{P}) - D(x_0, y_0|P)\right),$$

then it would be trivial to find values a_∂ and b_∂ such that

$$\Pr\left(-b_\alpha \leq D(x_0,y_0|\hat{P}) - D(x_0,y_0|P) \leq -a_\alpha\right) = 1-\alpha. \tag{10.27}$$

Of course, a_α and b_α are unknown, but from the empirical bootstrap distribution of the pseudoestimates $D_b(x_0,y_0|\hat{P}^*)$, $b=1,\ldots,B$, we can find values of $\hat{a}_\partial$ and $\hat{b}_\alpha$ such that

$$\Pr\left(-\hat{b}_\partial \leq D(x_0,y_0|\hat{P}^*) - D(x_0,y_0|\hat{P}) \leq -a_\partial|\hat{F}(S_n)\right) \approx 1-\partial. \tag{10.28}$$

Finding $\hat{a}_\alpha$ and $\hat{b}_\alpha$ involves sorting the values

$$\left(D_b(x_0,y_0|\hat{P}^*) - D(x_0,y_0|\hat{P})\right), \quad b=1,\ldots,B,$$

in increasing order and then deleting $\left(\frac{\alpha}{2}\times 100\right)$-percent of the elements at either end of the sorted list. Then set $-\hat{b}_\alpha$ and $-\hat{a}_\alpha$ equal to the endpoints of the truncated, sorted array, with $\hat{a}_\alpha \leq \hat{b}_\alpha$.

The bootstrap approximation of (10.27) is then

$$\Pr\left(-\hat{b}_\alpha \leq D(x_0,y_0|\hat{P}) - D(x_0,y_0|P) \leq -\hat{a}_\alpha\right) \approx 1-\alpha. \tag{10.29}$$

The estimated $(1-\alpha)$-percent confidence interval is then

$$D(x_0,y_0|\hat{P}) + \hat{a}_\alpha \leq D(x_0,y_0|P) \leq D(x_0,y_0|\hat{P}) + \hat{b}_\alpha. \tag{10.30}$$

This procedure can be used for any $(x_0,y_0) \in R_+^{p+q}$ for which $D(x_0,y_0|\hat{P}^*)$ exists. Typically, the applied researcher is interested in the efficiency scores of the observed units themselves; in this case, the above procedure can be repeated n times, with $(x_0,y_0)=(x_i,y_i)$, $i=1,\ldots,n$, producing a set of n confidence intervals of the form (10.30), one for each firm.

The method described above for estimating confidence intervals from the set of bootstrap values $\left\{D_b(x_0,y_0|\hat{P}^*)\right\}_{b=1}^B$ differs slightly from what we proposed in Simar and Wilson (1998); here, we avoid explicit use of a bias estimate, which adds unnecessary noise to the estimated confidence intervals. The bias estimate itself, however, remains interesting.

By definition,

$$\text{BIAS}\left(D(x_0,y_0|\hat{P})\right) = E\left(D(x_0,y_0|\hat{P})\right) - D(x_0,y_0|P). \tag{10.31}$$

The bootstrap bias estimate for the original estimator $D(x_0, y_0|\hat{P})$ is the empirical analog of (10.31):

$$\widehat{\text{BIAS}}_B\left(D(x_0, y_0|\hat{P})\right) = B^{-1}\sum_{b=1}^{B} D_b(x_0, y_0|\hat{P}^*) - D(x_0, y_0|\hat{P}). \tag{10.32}$$

It is tempting to construct a bias-corrected estimator of $D(x_0, y_0|P)$ by computing

$$\begin{aligned}\hat{\hat{D}}(x_0, y_0) &= D(x_0, y_0|\hat{P}) - \widehat{\text{BIAS}}_B\left(D(x_0, y_0|\hat{P})\right)\\ &= 2D(x_0, y_0|\hat{P}) - B^{-1}\sum_{b=1}^{B} D_b(x_0, y_0|\hat{P}^*).\end{aligned} \tag{10.33}$$

It is well known (e.g., Efron and Tibshirani 1993), however, that this bias correction introduces additional noise; the mean-square error of $\hat{\hat{D}}(x_0, y_0)$ may be greater than the mean-square error of $D(x_0, y_0|\hat{P})$. Thus the bias-correction in (10.33) should be used with caution.[5]

10.7 Implementing the Bootstrap

We have shown elsewhere (Kneip et al. 2008, 2010; Simar and Wilson 1998, 1999a, b, 2000a, b, 2008, 2009) that the generation of the bootstrap pseudodata S_n^* is crucial in determining whether the bootstrap gives consistent estimates of confidence intervals, bias, etc. In the classical linear regression model, one may resample from the estimated residuals, or alternatively resample the original sample observations to construct the pseudodataset S_n^*; in either case, the bootstrap will give consistent estimates. Both these approaches are variants of what has been called the *naive* bootstrap. The analog of these methods in frontier estimation would be to resample from the original distance function estimates, or alternatively from the (x, y) pairs in S_n.

In the present setting, however, there is a crucial difference from the classical linear regression model: here, the DGP F has bounded support over P, while in the classical linear regression model, the DGP has unbounded support. A related problem is that under our assumptions, the conditional density $f(D(x, y|P)$ $|x, \eta)$. has bounded support over the interval $(0,\ 1]$, and is rig ht-discontinuous at 1. It is widely known that problems such as these can cause the naive bootstrap to

[5] The mean-square error of the bias-corrected estimator in (10.33) could be evaluated in a second-level bootstrap along the same lines as the iterated bootstrap we propose below in Sect. 10.9. See Efron and Tibshirani (1993, pp. 138) for a simple example in a different context.

give *inconsistent* estimates (e.g., see Bickel and Freedman 1981; Swanepoel 1986; Beran and Ducharme 1991; Efron and Tibshirani 1993), and this is true in the present setting as we have shown in Simar and Wilson (1999a, b, 2000a).[6]

To address the boundary problems which doom the naive bootstrap in the present situation, we draw pseudodatasets from a smooth, consistent, nonparametric estimate of the DGP F, as represented by $f(x, \omega, \eta)$ in (10.15). In Simar and Wilson (1998), we drew values D^* from a kernel estimate $\hat{f}(D)$ of the marginal density of the original estimates $D(x_i, y_i|\hat{P})$, $i = 1, \ldots, n$; given (10.16), this is tantamount to assuming $f(\omega|x, \eta) = f(\omega)$ in (10.15), or in other words, that the distribution of inefficiencies is homogeneous and does not depend upon location within the production set P. This assumption can be relaxed, as in Simar and Wilson (2000b), but at a cost of increased complexity and computational burden. In this chapter, we focus on the homogeneous case. In applications, one can test the homogeneity assumption using the methods surveyed by Wilson (2003), and then choose between the bootstrap methods in Simar and Wilson (1998, 2000b, 2009), Kneip et al. (2010). Note that for the FDH estimator, only the bootstrap by subsampling has been suggested and its consistency proved in Jeong and Simar (2006).

Kernel density estimation is rather easy to apply and has been widely studied. Given values z_i, $i = 1, \ldots, n$ on the real number line, the kernel estimate of the density $g(z)$ is given by

$$\hat{g}(z) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{z - z_i}{h}\right), \tag{10.34}$$

where $K(\cdot)$ is a *kernel function* and h is a *bandwidth parameter*. Both $K(\cdot)$ and h must be chosen, but there are well-established guidelines to aid with these choices. In particular, the kernel function $K(\cdot)$ must be piecewise continuous and must satisfy $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$ and $\int_{-\infty}^{\infty} uK(u)\mathrm{d}u = 0$. Thus, any probability density function that is symmetric around zero is a valid kernel function; in addition, one could use even-ordered polynomials bounded on some interval with coefficients chosen so that the polynomial integrates to unity over this interval.[7] As Silverman (1986) shows, the choice of the kernel function is far less critical for obtaining a good estimate (in the sense of low mean integrated square error) of $f(z)$ than the choice of the bandwidth parameter h.

[6] Explicit descriptions of why either variation of the naive bootstrap results in inconsistent estimates are given in Simar and Wilson (1999a, 2000a). Löthgren and Tambour (1997, 1999) and Löthgren (1998, 1999) employ a bizarre, illogical variant of the naive bootstrap different from the more typical variations we have mentioned. This approach also leads to an inconsistency problem, as discussed and confirmed with Monte Carlo experiments in Simar and Wilson (2000a).

[7] Appropriately chosen high-order kernels can reduce the order of the bias in the kernel density estimator, but run the risk of producing negative density estimates at some locations.

In order for the kernel density estimator to be consistent, the bandwidth must be chosen so that $h = O(n^{-1/5})$ for the univariate case considered here. If the data are approximately normally distributed, then one may employ the normal reference rule and set $h = 1.06\,\hat{\sigma}\,n^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation of the data whose density is being estimated (Silverman 1986). In cases where the data are clearly non-normal, as will be the case when estimating the density $f(D)$, one can plot the density estimate for various values of h and choose the value of h that gives a reasonable estimate, as in Silverman (1978), Simar and Wilson (1998). This approach, however, contains an element of subjectivity; a better approach is to employ least squares cross-validation, which involves choosing the value of h that minimizes an approximation to mean integrated square error; see Silverman (1986) for details. In many cases involving high dimensions (i.e., large $p + q$), many of the distance function estimates will equal unity (this is especially the case when the convex hull or FDH estimators are used), and this creates a discretization problem in the cross-validation procedure. Simar and Wilson (2002) proposed a weighted cross-validation procedure to avoid this problem in the univariate setting; Simar and Wilson (2000b) extend this idea to the multivariate setting.

Regardless of how the bandwidth parameter is chosen, it is important to note that ordinary kernel estimators (such as the one in (10.34) above) of densities with bounded support will be biased near the boundaries of the support. Necessarily, for $(x, y) \in P$, we have $0 < D(x, y|P) \leq D(x, y|\hat{P}) \leq 1$, and so this will be a problem in estimating $f(D)$. It is easy to see why this problem arises: in (10.34), when $K(\cdot)$ is symmetric about zero, the density estimate is determined at a particular point on the real number line by adding up the value of functions $K(\cdot)$ centered over the observed data along the real number line *on either side* of the point where the density estimate is being evaluated. If, for example, $K(\cdot)$ is chosen as a standard normal probability density function, then nearby observations contribute relatively more to the density estimate at this particular point than do observations that are farther away along the real number line.[8] When (10.34) is used to evaluate the density estimate at unity in our application, then if $K(\cdot)$ is symmetric, there will necessarily be no data on the right side of the boundary to contribute to the smoothing, and this causes the bias problem. An obvious approach would be to let the kernel function become increasingly asymmetric as the boundary is approached, but this is problematic for several reasons as discussed by Scott (1992). A much simpler solution is to use the reflection method as in Simar and Wilson (1998).

[8] This is the sense in which the kernel density estimator is a *smoother,* since it is, in effect, smoothing the empirical density function which places probability mass $1/n$ at each observed datum. Setting $h = 0$ in (10.34) yields the empirical density function, while letting $h \to \infty$ yields a flat density estimate. The requirement that $h = O(n^{-1/5})$ to ensure consistency of the kernel density estimate results from the fact that as n increases, h must become smaller, but not too quickly.

The reflection method involves reflecting each of the n original estimates $D(x_i, y_i|\hat{P})$ about the boundary at unity (by computing $1 - D(x_i, y_i|\hat{P})$ for each $D(x_i, y_i|\hat{P})$, $i = 1, \ldots, n$) to obtain $2n$ points along the real number line. The output distance function estimates are also bounded on the left at 0, but typically there will be no values near zero, suggesting that the density is near zero at this boundary. Hence, we ignore the effect of the left boundary, and concentrate on the boundary at unity. Viewing the reflected data (with $2n$observations) as unbounded, we can estimate the density of these data using the estimator in (10.34) with no special problems. Then, this unbounded density estimate can be truncated on the right at unity to obtain an estimate of the density of the $D(x_i, y_i|\hat{P})$ with support on the interval $(0, 1]$.

Once the kernel function $K(\cdot)$ and the bandwidth h have been chosen, it is not necessary to evaluate the kernel density estimate in (10.34) in order to draw random values. Rather, a computational shortcut is afforded by Silverman (1986) and for cases where the kernel function $K(\cdot)$ is a regular probability density function.

We need to draw h values D^* from the estimated density of the original distance function estimates. Let $\{\varepsilon_i\}_{i=1}^n$ be a set of n iid draws from the probability density function used to define the kernel function; let $\{d_i\}_{i=1}^n$ be a set of values drawn independently, uniformly, and with replacement from the set of reflected distance function estimates $R = \left\{D(x_i, y_i|\hat{P}), 2 - D(x_i, y_i|\hat{P})\right\}$; and let $\bar{d} = n^{-1}\sum_{i=1}^n d_i$. Then compute

$$d_i^* = \bar{d} + (1 + h^2/s^2)^{-1/2}(d_i + h\varepsilon_i - \bar{d}), \tag{10.35}$$

where s^2 is the sample variance of the values $v_i = d_i + h\varepsilon_i$. Using the convolution theorem, it can be shown that $v_i \sim \hat{g}(\cdot)$, where $\hat{g}(\cdot)$ is the kernel estimate of the density of the original distance function estimates and their reflections in R. As is typical with kernel density estimates, the variance of the v_i must be scaled upward as in (10.35). Straightforward manipulations reveal that

$$E(d_i^*|R) = 1, \tag{10.36}$$

and

$$VAR(d_i^*|R) = s^2\left(1 + \frac{h^2}{n(s^2 + h^2)}\right), \tag{10.37}$$

so that the variance of the d_i^* is asymptotically correct. All that remains is to reflect the d_i^* about unity by computing, for each $i = 1, \ldots, n$,

$$D_i^* = \begin{cases} d_i^* & \text{if } d_i^* \leq 1; \text{ or} \\ 2 - d_i^* & \text{otherwise.} \end{cases} \tag{10.38}$$

The last step is equivalent to "folding" the right half of the symmetric (about 1) estimate $\hat{g}(\cdot)$ to the left around 1, and ensures that $d_i^* \leq 1$ for all i. Putting all of this together, bootstrap estimates of confidence intervals, bias, etc., for the output distance function $D(x_0, y_0|P)$ evaluated for a particular, arbitrary point $(x_0, y_0) \in R_+^{p+q}$ can be obtained via the following algorithm:

Algorithm #1.

[1] For each $(x_i, y_i) \in S_n$, apply one of the distance function estimators in (10.11)–(10.13) to obtain estimates $D(x_i, y_i|\hat{P})$, $i = 1, \ldots, n$.
[2] If $(x_0, y_0) \notin S_n$, repeat step [1] for (x_0, y_0) to obtain $D(x_0, y_0|\hat{P})$.
[3] Reflect the n estimates $D(x_i, y_i|\hat{P})$ about unity, and determine the bandwidth parameter h via least-squares cross-validation.
[4] Use the computational shortcut in (10.35) to draw n bootstrap values D_i^*, $i = 1, \ldots, n$, from the kernel density estimate of the efficiency estimates from step [1] and their reflected values from step [3].
[5] Construct a pseudodataset S_n^* with elements (x_i^*, y_i^*) given by $y_i^* = D_i^* y_i / D \times (x_i, y_i|\hat{P})$ and $x_i^* = x$.
[6] Use (10.25) (or an analog of (10.25) if the convex cone estimators were used in step [1]) to compute the bootstrap estimate $D^*(x_0, y_0|\hat{P}^*)$, where $\hat{P}^*$ denotes the convex hull of the FDH of the bootstrap sample S_n^*.
[7] Repeat steps [4]–[6] B times to obtain a set of B bootstrap estimates

$$\left\{D_b(x_0, y_0|\hat{P}^*)\right\}_{b=1}^{B}.$$

[8] Use (10.28) to determine $\hat{a}_\partial$ and $\hat{b}_\partial$, and then use these in (10.30) together with the original estimate $D^*(x_0, y_0|\hat{P}^*)$ obtained in step [2] (or step [1] if $(x_i, y_i) \in S_n$) to obtain an estimated confidence interval for $D(x_0, y_0|\hat{P}^*)$. In addition, the bootstrap estimates can be used to obtain an estimate of the bias of $D(x_0, y_0|P)$ from (10.32), and to obtain the bias-corrected estimator in (10.33) if desired.

Note that the arbitrary point (x_0, y_0) might or might not be in P. Typically, however, researchers are interested in the efficiency of firms represented in the observed sample, S_n. In that case, (x_0, y_0) will correspond in turn with each of the $(x_i, y_i) \in S_n$. Rather than repeatedly apply Algorithm #1 n times, considerable computational cost can be saved by noting that step [2] is not applicable in this case, and performing step [6] n times (for each element of S_n) inside each of the bootstrap loops in step [7]. This will result in n distinct sets of bootstrap estimates when step [8] is reached; each can then be used to estimate a confidence interval for its corresponding true distance function.

It is rather straightforward to extend Algorithm #1 to estimation of confidence intervals for Malmquist indices (Simar and Wilson 1999c) or to the problem of testing hypotheses regarding returns to scale as in Simar and Wilson (2002). In choosing

between the convex hull and convex cone estimators defined in (10.11)–(10.12), it is important to have some idea of whether the true technology P^{∂} exhibits constant returns to scale everywhere; if it does not, the convex cone estimator in (10.12) will necessarily be inconsistent. It would be trivial to extend the results in Simar and Wilson (2002) to test for concavity of the production set P using the convex hull estimator and the FDH estimator defined in (10.13). Tests of other hypotheses regarding the shape of the technology should also be possible.

10.8 Monte Carlo Evidence

We conducted a series of Monte Carlo experiments to examine the coverage probabilities of estimated confidence intervals for output distance functions. We simulated two different DGPs for these Monte Carlo experiments, but to minimize computational costs, we set $p = q = 1$ in both cases so that we examine only a single output produced from a single input. To consider a true technology with constant returns to scale everywhere, we used the model

$$y = x\mathrm{e}^{-|v|}, \tag{10.39}$$

where $v \sim N(0,1)$ and $x \sim$ Uniform on (1,9). To allow for variable returns to scale, we used the model

$$y = (x-2)^{2/3}\mathrm{e}^{-|v|}, \tag{10.40}$$

in experiments where this model was used, $v \sim N(0,1)$ as in the previous case, but $x \sim$ Uniform on (3,12). Pseudorandom uniform deviates were generated from a multiplicative congruential pseudorandom number generator with modulus $2^{31}-1$ and multiplier 7^5 (see Lewis et al. 1969). Pseudorandom normal deviates were generated via the Box–Muller method (see Press et al. 1986, pp. 202–203). In each Monte Carlo trial, the fixed point is given by $x_0 = 7.5$, $y_0 = 2$.

Each Monte Carlo experiment involved 1,000 Monte Carlo trials; on each trial, we performed $B = 2,000$ bootstrap replications. For each trial, we estimated confidence intervals for $D(x_0, y_0|P)$ and then checked whether the estimated confidence intervals included the true value $D(x_0, y_0|P)$. After 1,000 trials, we divided the number of cases where $D(x_0, y_0|P)$ was included in the estimated confidence intervals by 1,000 to produce the results shown in Table 10.1.

Table 10.1 contains six numbered columns; the first three correspond to the constant returns to scale model in (10.39) where the convex cone estimator defined in (10.12) was employed, while the last three correspond to the variable returns to scale model in (10.40) where the convex hull estimator defined in (10.11) was used. The columns labeled "Smooth" correspond to Algorithm #1 given earlier, using the kernel smoothing described in Sect. 10.7. Results in the columns labeled "(x, y)" were obtained using the naive bootstrap where resampling is from the elements of S_n.

Table 10.1 Monte Carlo estimates of confidence interval coverages ($p = q = 1$; i.e., one input, one output)

n	Significance Level	(1) Smooth	(2) x, y	(3) $D(x, y\|\hat{P})$	(4) Smooth	(5) x, y	(6) $D(x, y\|\hat{P})$
10	0.80	0.745	0.757	0.757	0.626	0.757	0.450
	0.90	0.858	0.789	0.789	0.754	0.830	0.609
	0.95	0.916	0.899	0.899	0.852	0.917	0.740
	0.975	0.948	0.936	0.936	0.895	0.938	0.826
	0.99	0.976	0.957	0.957	0.941	0.969	0.894
25	0.80	0.750	0.770	0.771	0.646	0.683	0.393
	0.90	0.861	0.776	0.775	0.777	0.808	0.562
	0.95	0.932	0.894	0.890	0.843	0.889	0.689
	0.975	0.965	0.928	0.922	0.905	0.926	0.807
	0.99	0.980	0.967	0.950	0.944	0.963	0.885
50	0.80	0.752	0.769	0.765	0.645	0.683	0.369
	0.90	0.863	0.783	0.778	0.770	0.808	0.515
	0.95	0.920	0.896	0.891	0.842	0.889	0.640
	0.975	0.950	0.934	0.941	0.901	0.926	0.751
	0.99	0.972	0.962	0.960	0.947	0.963	0.850
100	0.80	0.766	0.767	0.749	0.620	0.580	0.285
	0.90	0.864	0.797	0.800	0.749	0.724	0.438
	0.95	0.921	0.889	0.891	0.830	0.810	0.565
	0.975	0.956	0.929	0.947	0.882	0.869	0.676
	0.99	0.980	0.955	0.970	0.925	0.927	0.786
200	0.80	0.780	0.757	0.736	0.600	0.531	0.280
	0.90	0.877	0.795	0.803	0.728	0.674	0.406
	0.95	0.937	0.879	0.888	0.811	0.774	0.514
	0.975	0.963	0.939	0.939	0.867	0.825	0.607
	0.99	0.984	0.961	0.963	0.902	0.886	0.726
400	0.80	0.785	0.759	0.772	0.632	0.540	0.306
	0.90	0.878	0.809	0.811	0.758	0.665	0.430
	0.95	0.936	0.883	0.889	0.845	0.754	0.529
	0.975	0.970	0.937	0.936	0.891	0.828	0.626
	0.99	0.990	0.963	0.961	0.926	0.896	0.720
800	0.80	0.767	0.769	0.733	0.614	0.487	0.281
	0.90	0.884	0.811	0.791	0.730	0.616	0.380
	0.95	0.950	0.886	0.871	0.810	0.708	0.478
	0.975	0.971	0.938	0.930	0.858	0.790	0.571
	0.99	0.984	0.966	0.955	0.907	0.853	0.667
1,600	0.80	0.786	0.752	0.735	0.663	0.522	0.302
	0.90	0.897	0.799	0.793	0.777	0.657	0.423
	0.95	0.957	0.876	0.868	0.846	0.739	0.517
	0.975	0.981	0.945	0.929	0.896	0.815	0.629
	0.99	0.990	0.970	0.960	0.944	0.887	0.730
3,200	0.80	0.801	0.758	0.765	0.653	0.500	0.303
	0.90	0.904	0.822	0.813	0.775	0.639	0.430

(continued)

Table 10.1 (continued)

n	Significance Level	(1) Smooth	(2) x, y	(3) $D(x,y\|\hat{P})$	(4) Smooth	(5) x, y	(6) $D(x,y\|\hat{P})$
	0.95	0.951	0.897	0.864	0.860	0.717	0.527
	0.975	0.976	0.951	0.924	0.899	0.785	0.611
	0.99	0.992	0.973	0.951	0.938	0.865	0.715
6,400	0.80	0.839	0.765	0.743	0.666	0.537	0.337
	0.90	0.917	0.816	0.811	0.789	0.656	0.462
	0.95	0.960	0.878	0.868	0.868	0.737	0.588
	0.975	0.977	0.943	0.937	0.909	0.806	0.671
	0.99	0.987	0.964	0.962	0.942	0.873	0.744

In terms of Algorithm #1, this amounts to deleting step [3] and replacing steps [4]–[5] with a single step where we draw n observations from S_n uniformly, independently, and with replacement to form the pseudosample S_n^*. Results in the columns labeled "$D(x,y|\hat{P})$" were obtained with the other variant of the naive bootstrap mentioned previously, i.e., where the original distance function estimates were resampled. Again in terms of Algorithm #1, this involved deleting step [3] and replacing step [4] with a step where we draw n times uniformly, independently, and with replacement, from the set $\left\{D(x_i,y_i|\hat{P})\right\}_{i=1}^{n}$ to obtain the bootstrap values D_i^*.

Under constant returns to scale, the smooth bootstrap performs reasonably well with only ten observations in the single input, single output case. The minor fluctuations in the reported coverages in column (1) that are seen as the sample size increases beyond 25 are to be expected due to sampling variation in the Monte Carlo experiment, and due to the fact that a finite number of bootstrap replications are being used. Also, the two variants of the naive bootstrap appear to perform, for the particular model in (10.39), reasonably well in terms of coverages, although not as well as the smooth bootstrap. One should not conclude from this, however, that it is safe to use the naive bootstrap – since it is inconsistent, there is no reason to think that coverages would be reasonable in any other setting than the one considered here. Moreover, the case of $p = q = 1$ with constant return to scale, although not typical of actual applications, is the most favorable case for the naive bootstrap since typically the estimated frontier will intersect only one sample observation. In higher dimensions, or where the convex hull or FDH estimators are used, a far higher portion of the original observations will have corresponding efficiency estimates equal to unity, which will be problematic for either version of the naive bootstrap.

Differences between the smooth bootstrap and the variants of the naive bootstrap become more pronounced with variable returns to scale, as columns (4)–(6) in Table 10.1 reveal. The smooth bootstrap does not perform quite as well as in the constant returns to scale case, but this is to be expected given the greater degree of curvature of the frontier in (10.40) (see Gijbels et al. 1999, for details on the relation between the curvature of the frontier and the convergence rate of the output distance function estimator used here). Also, the coverages obtained with the smooth bootstrap appear to fluctuate somewhat as the sample size increases, but eventually seem to be on a path toward the nominal significance levels. This merely reflects the fact that consistency does not imply that convergence must be monotonic. With the two variants of the naive bootstrap, however, even with 6,400 observations, coverages at all

Table 10.2 Monte Carlo estimates of confidence interval coverages showing the influence of range of the input for the variable returns to scale case ($p = q = 1$, $n = 200$)

Significance	$x \sim U(6.5, 8.5)$	$x \sim U(5.5, 9.5)$	$x \sim U(4.5, 10.5)$	$x \sim U(3.5, 11.5)$
Smooth bootstrap				
0.80	0.734	0.727	0.659	0.633
0.90	0.846	0.829	0.790	0.748
0.95	0.906	0.888	0.858	0.830
0.975	0.946	0.930	0.902	0.886
0.99	0.967	0.963	0.946	0.920
Resample (x, y)				
0.80	0.686	0.620	0.565	0.540
0.90	0.813	0.763	0.719	0.689
0.95	0.897	0.849	0.808	0.769
0.975	0.941	0.913	0.872	0.839
0.99	0.969	0.955	0.932	0.902
Resample $D(x, y\|\hat{P})$				
0.80	0.326	0.276	0.274	0.279
0.90	0.489	0.442	0.406	0.398
0.95	0.640	0.568	0.526	0.516
0.975	0.744	0.686	0.635	0.615
0.99	0.858	0.801	0.753	0.725

significance levels are worse than with the smooth bootstrap at any of the sample sizes we examined. Again, the problem is that with variable returns to scale, many sample observations will typically have efficiency estimates equal to unity. Between the two variants of the naive bootstrap, the one based on resampling from the original distance function estimates seems to give poorer coverage than the one based on resampling from S_n; but this distinction is irrelevant, since both methods yield inconsistent estimates.

It is important to note that the performance of the bootstrap will vary not only as sample size increases, but also depending on features of the true model, particularly the curvature of the frontier and the shape of the density $f(x, y)$, The DGP represented by (10.40) is somewhat unfavorable to the smooth bootstrap, in that the range of the inputs is considerably larger than the range of the outputs. To illustrate this point, we performed additional Monte Carlo experiments for $n = 200$, generating data for each trial using (10.40), but varying the range of the input variable. Results from these experiments are shown in Table 10.2, where we report estimated confidence interval coverages as in Table 10.1. We considered four cases, where x is alternately distributed uniformly on the (6.5, 8.5), (5.5, 9.5), (4.5, 10.5), and (3.5, 11.5) intervals (other features of the experiments remained unchanged). The results show that as the distribution of the input becomes more disperse, the coverages of confidence intervals estimated with each method worsen. As before, however, the smooth bootstrap dominates both variants of the naive bootstrap in every instance.

Table 10.3 Widths of estimated confidence intervals for constant returns to scale case

n	Mean	Standard deviation	Range
Smooth bootstrap			
10	0.1384	0.0335	0.0485–0.2838
25	0.0551	0.0099	0.0271–0.0935
50	0.0283	0.0044	0.0183–0.0477
100	0.0146	0.0019	0.0094–0.0255
200	0.0076	0.0008	0.0055–0.0110
400	0.0039	0.0004	0.0029–0.0053
800	0.0019	0.0002	0.0016–0.0026
1,600	0.0010	0.0001	0.0008–0.0012
3,200	0.0005	0.0001	0.0004–0.0007
6,400	0.0003	0.0001	0.0002–0.0003
Resample (x, y)			
10	0.2018	0.1404	0.0079–0.9945
25	0.0664	0.0384	0.0048–0.2556
50	0.0320	0.0186	0.0018–0.1391
100	0.0154	0.0087	0.0011–0.0527
200	0.0078	0.0046	0.0001–0.0306
400	0.0037	0.0021	0.0001–0.0141
800	0.0019	0.0011	0.0002–0.0066
1,600	0.0009	0.0005	0.0–0.0040
3,200	0.0005	0.0003	0.0–0.0018
6,400	0.0002	0.0001	0.0–0.0009
Resample $D(x, y\|\hat{P})$			
10	0.2018	0.1404	0.0079–0.9945
25	0.0693	0.0431	0.0038–0.3975
50	0.0315	0.0175	0.0018–0.1218
100	0.0157	0.0089	0.0011–0.0580
200	0.0074	0.0040	0.0003–0.0295
400	0.0038	0.0022	0.0002–0.0141
800	0.0019	0.0010	0.0001–0.0072
1,600	0.0009	0.0006	0.0001–0.0034
3,200	0.0005	0.0003	0.0–0.0017
6,400	0.0002	0.0001	0.0–0.0009

Coverage probabilities of estimated confidence intervals tell only part of the story. Consequently, we also examined the widths of the confidence intervals estimated in the experiments reported in Table 10.1. Table 10.3 shows, for the case of constant returns to scale, the mean width of the estimated confidence intervals over 1,000 Monte Carlo trials, the standard deviation of these widths, and the range of the widths of the estimated confidence intervals. Up to about 100 observations, the smooth bootstrap produces narrower confidence interval estimates than either of the naive bootstrap methods. Differences beyond $n = 100$ appear inconsequential. Table 10.3 shows similar results for the case of variable returns to

Table 10.4 Widths of estimated confidence intervals for variable returns to scale case

n	Mean	Standard deviation	Range
Smooth bootstrap			
10	0.3020	0.1520	0.0499–2.3832
25	0.1331	0.0407	0.0426–0.3941
50	0.0741	0.0195	0.0208–0.1619
100	0.0426	0.0103	0.0181–0.0819
200	0.0254	0.0053	0.0133–0.0460
400	0.0161	0.0030	0.0081–0.0271
800	0.0102	0.0018	0.0057–0.0167
1,600	0.0066	0.0011	0.0036–0.0105
3,200	0.0042	0.0007	0.0022–0.0062
6,400	0.0028	0.0004	0.0016–0.0043
Resample (x, y)			
10	0.4911	0.2834	0.0603–2.9616
25	0.1667	0.0741	0.0237–0.4875
50	0.1667	0.0741	0.0237–0.4875
100	0.0447	0.0185	0.0101–0.1252
200	0.0245	0.0098	0.0046–0.0723
400	0.0147	0.0060	0.0035–0.0417
800	0.0089	0.0034	0.0021–0.0262
1,600	0.0057	0.0022	0.0012–0.0153
3,200	0.0035	0.0013	0.0007–0.0088
6,400	0.0022	0.0009	0.0002–0.0059
Resample $D(x, y\|\hat{P})$			
10	0.2701	0.1746	0.0264–1.6783
25	0.1067	0.0529	0.0169–0.4355
50	0.0529	0.0241	0.0039–0.1742
100	0.0270	0.0125	0.0057–0.0800
200	0.0147	0.0064	0.0014–0.0392
400	0.0090	0.0037	0.0015–0.0229
800	0.0054	0.0021	0.0011–0.0123
1,600	0.0036	0.0013	0.0003–0.0092
3,200	0.0023	0.0008	0.0005–0.0053
6,400	0.0015	0.0005	0.0003–0.0031

scale. Here, the differences between the smooth bootstrap and the naive variants are more pronounced at large sample sizes. The smooth bootstrap yields narrower intervals than the naive methods at smaller sample sizes, but wider intervals at very large sample sizes. With either constant or variable returns to scale, the widths of the confidence interval estimates from the smooth bootstrap have smaller standard deviations than those from the naive methods.

In Tables 10.4 and 10.5, we give results on the bias estimates (based on (10.32)) for the constant returns to scale case (Table 10.4) and the variable returns to scale

Table 10.5 Bootstrap bias estimates constant returns to scale case

n	"True"	Mean	Standard deviation	Range
Smooth bootstrap				
10	0.0517	0.0362	0.0085	0.0133–0.0756
25	0.0203	0.0147	0.0025	0.0077–0.0241
50	0.0101	0.0076	0.0011	0.0050–0.0123
100	0.0048	0.0039	0.0005	0.0025–0.0067
200	0.0024	0.0020	0.0002	0.0015–0.0029
400	0.0012	0.0010	0.0001	0.0008–0.0014
800	0.0006	0.0005	0.0001	0.0004–0.0007
1,600	0.0003	0.0003	0.0000	0.0002–0.0004
3,200	0.0002	0.0001	0.0000	0.0001–0.0002
6,400	0.0001	0.0001	0.0	0.0001–0.0001
Resample (x, y)				
10	0.0517	0.0324	0.0250	0.0025–0.2503
25	0.0203	0.0117	0.0082	0.0008–0.0585
50	0.0101	0.0058	0.0039	0.0003–0.0278
100	0.0048	0.0028	0.0019	0.0003–0.0152
200	0.0024	0.0014	0.0010	0.0001–0.0066
400	0.0012	0.0007	0.0005	0.0–0.0034
800	0.0006	0.0004	0.0003	0.0–0.0017
1,600	0.0003	0.0002	0.0001	0.0–0.0009
3,200	0.0002	0.0001	0.0001	0.0–0.0005
6,400	0.0001	0.0000	0.0001	0.0–0.0003
Resample $D(x, y\|\hat{P})$				
10	0.0517	0.0324	0.0250	0.0025–0.2503
25	0.0203	0.0121	0.0085	0.0008–0.0650
50	0.0101	0.0057	0.0037	0.0003–0.0302
100	0.0048	0.0030	0.0020	0.0003–0.0135
200	0.0024	0.0014	0.0009	0.0001–0.0071
400	0.0012	0.0007	0.0005	0.0001–0.0036
800	0.0006	0.0004	0.0002	0.0–0.0016
1,600	0.0003	0.0002	0.0001	0.0–0.0009
3,200	0.0002	0.0001	0.0001	0.0–0.0006
6,400	0.0001	0.0000	0.0000	0.0–0.0002

case (Table 10.6). In both tables, the first column gives the sample size, while the second column (labeled "True") gives the Monte Carlo estimate of the true bias, i.e., the mean of the original efficiency estimates $D(x_0, y_0|\hat{P})$ produced in step [1] of Algorithm #1 over the 1,000 Monte Carlo Trials, minus the known, true value $D(x_0, y_0|P) = 0.6419$. The remaining columns in the tables give the mean and standard deviation of the bias estimates (10.32) over 1,000 Monte Carlo trials, as well as the range of the bias estimates.

In both the constant and variable returns to scale cases, the naive methods yield, on average, smaller and less accurate bias estimates than the smooth bootstrap, with

Table 10.6 Bootstrap bias estimates variable returns to scale case

n	"True"	Mean	Standard deviation	Range
Smooth bootstrap				
10	0.1615	0.1017	0.0487	0.0149–0.7556
25	0.0735	0.0511	0.0171	0.0140–0.1452
50	0.0403	0.0298	0.0090	0.0074–0.0745
100	0.0247	0.0178	0.0050	0.0066–0.0403
200	0.0150	0.0110	0.0029	0.0049–0.0216
400	0.0088	0.0071	0.0017	0.0031–0.0132
800	0.0060	0.0045	0.0010	0.0022–0.0090
1,600	0.0035	0.0029	0.0007	0.0013–0.0054
3,200	0.0023	0.0019	0.0004	0.0009–0.0032
6,400	0.0014	0.0013	0.0003	0.0007–0.0021
Resample (x, y)				
10	0.1615	0.2263	0.5805	−3.3019–0.3282
25	0.0735	0.0337	0.0320	−0.5655–0.1450
50	0.0735	0.0337	0.0320	−0.5655–0.1450
100	0.0247	0.0100	0.0048	0.0015–0.0346
200	0.0150	0.0057	0.0028	0.0013–0.0183
400	0.0088	0.0034	0.0016	0.0006–0.0123
800	0.0060	0.0020	0.0010	0.0004–0.0064
1,600	0.0035	0.0013	0.0006	0.0002–0.0037
3,200	0.0023	0.0008	0.0004	0.0001–0.0021
6,400	0.0013	0.0005	0.0002	0.0001–0.0017
Resample $D(x, y\|\hat{P})$				
10	0.1615	0.0506	0.0412	0.0030–0.6932
25	0.0735	0.0207	0.0142	0.0017–0.1226
50	0.0403	0.0101	0.0069	0.0005–0.0642
100	0.0247	0.0052	0.0035	0.0004–0.0224
200	0.0150	0.0029	0.0020	0.0002–0.0126
400	0.0088	0.0018	0.0012	0.0001–0.0076
800	0.0060	0.0011	0.0007	0.0001–0.0044
1,600	0.0035	0.0008	0.0005	0.0–0.0030
3,200	0.0023	0.0005	0.0003	0.0001–0.0016
6,400	0.0013	0.0003	0.0002	0.0–0.0010

the differences becoming more pronounced in the variable returns to scale case. With variable returns to scale, the naive methods typically give bias estimates whose averages are equal to about half the average of the corresponding bias estimates from the smooth bootstrap. The naive methods also produce bias estimates with larger standard deviation than in the case of the smooth bootstrap for sample sizes up to 100, but these differences become minimal for larger sample sizes. The naive method based on resampling (x, y) pairs yields negative bias estimates in some cases for $n = 10, 25$, and 50 with variable returns to scale – even though bias is necessarily positive for our simulated DGP.

10.9 Enhancing the Performance of the Bootstrap

The results of the preceding section show that in less favorable situations, even if the bootstrap is consistent, the coverage probabilities could be poorly approximated in finite samples.

Let I_α denote the $(1-\alpha)$-level confidence interval for $D(x_0, y_0|P)$ given by (10.30). We have

$$I_\alpha = \left[D(x_0, y_0|\hat{P}) + \hat{a}_\alpha, D(x_0, y_0|\hat{P}) + \hat{b}_\alpha\right],$$

where $\hat{a}_\alpha$ and $\hat{b}_\alpha$ are solutions to (10.28). Our Monte Carlo experiments indicate that, in finite samples, the error $\Pr(D(x_0, y_0|P) \in I_\partial) - (1-\alpha)$ may sometimes be substantial. Of course, in practice, with real data, we cannot evaluate this error as in Monte Carlo experiments.

We show in this section that a method based on the iterated bootstrap may be used to estimate the true coverage,

$$\pi(\alpha) = \Pr(D(x_0, y_0|P) \in I_\alpha), \tag{10.41}$$

and also to estimate the value of α for which the true coverage is equal to a predetermined nominal level $(1-\alpha_0)$, such as 0.95. If $\hat{\pi}(\alpha)$ is our estimator of $\pi(\alpha)$, then we can search for a solution $\hat{\alpha}$ in the equation

$$\hat{\pi}(\hat{\alpha}) = 1 - \alpha_0, \tag{10.42}$$

and then recalibrate our initial confidence interval I_α, choosing instead $I_{\hat{\alpha}}$ as our final, corrected confidence interval for $D(x_0, y_0|P)$.

The idea of iterating the bootstrap is discussed in detail by Hall (1992), and has been used in parametric frontier models by Hall et al. (1995). Using the notation introduced above, the method may be defined as follows. After steps [5]–[6] of Algorithm #1, where S_n^* and $D(x_0, y_0\left|\hat{P}^*\right.)$ have been computed, we generate a second-level bootstrap sample S_n^{**} from S_n^* along the same lines by which we generated S_n^* from S_n.

In particular, the second level bootstrap estimator $D(x_0, y_0\left|\hat{P}^{**}\right.)$ corresponds to S_n^{**}; i.e., $\hat{P}^{**}$ is the convex hull of the FDH of the elements of S_n^{**}. So, from the bootstrap distribution of $D(x_0, y_0\left|\hat{P}^{**}\right.)$, we can find values $\hat{a}_\alpha^*$ and $\hat{b}_\alpha^*$ such that

$$\Pr\left(-\hat{b}^*_{\ \alpha} \le D(x_0, y_0\left|\hat{P}^{**}\right.) - D(x_0, y_0\left|\hat{P}^*\right.) \le -\hat{a}^*_\alpha|\hat{F}(S_n^*)\right) \approx 1 - \alpha, \tag{10.43}$$

where $D(x_0, y_0\left|\hat{P}^*\right.)$ and $\hat{F}(S_n^*)$ have replaced $D(x_0, y_0|\hat{P})$ and $\hat{F}(S_n)$ in (10.28).

This will provide a confidence interval for $D(x_0, y_0|\hat{P})$ (in the second-level bootstrap) denoted I^*_α, just as I_α was the confidence interval for $D(x_0, y_0|P)$ in the first-level bootstrap:

$$I^*_\alpha = \left[D(x_0, y_0\Big|\hat{P}^*) + \hat{a}^*_\alpha, D(x_0, y_0\Big|\hat{P}^*) + \hat{b}^*_\alpha\right]. \tag{10.44}$$

However, here $D(x_0, y_0|\hat{P})$ is known; so by repeating the first-level bootstrap many times, we can record $\hat{\pi}(\alpha)$, the proportion of times that I^*_α covers $D(x_0, y_0|\hat{P})$:

$$\hat{\pi}(\alpha) = \Pr\left(D(x_0, y_0|\hat{P}) \in I^*_\alpha\Big|\hat{F}(S_n)\right). \tag{10.45}$$

This quantity is the estimator of the function $\pi(\alpha)$ defined in (10.41); by solving (10.42), we obtain the iterated bootstrap confidence interval, $I_{\hat{\alpha}}$.

The algorithm can be implemented as follows:

Algorithm #2

[1] For each $(x_i, y_i) \in S_n$, apply one of the distance function estimators in (10.11)–(10.13) to obtain estimates $D(x_i, y_i|\hat{P})$, $i = 1, \ldots, n$. Here, the reference set is S_n.

[2] Repeat step [1] for (x_0, y_0) to obtain $D(x_0, y_0|\hat{P})$.

[3] Reflect the n estimates $D(x_i, y_i|\hat{P})$ about unity, and determine the bandwidth parameter h via least-squares cross-validation.

[4] Use the computational shortcut in (10.35) to draw n bootstrap values D^*_i, $i = 1, \ldots, n$, from the kernel estimate of the density of the $D(x_i, y_i|\hat{P})$.

[5] Construct a pseudodataset S^*_n with elements (x^*_i, y^*_i) given by $y^*_i = D^*_i y_i / D \times (x_i, y_i|\hat{P})$ and $x^*_i = x$.

[6] Use (10.25) (or an analog of (10.25) if the convex cone estimators were used in step [1]) to compute the bootstrap estimate $D(x_0, y_0|\hat{P}^*)$; here the reference set is S^*_n. [6.1] For each $(x^*_i, y^*_i) \in S^*_n$, apply the same distance function estimators chosen in [1] to obtain estimates $D(x^*_i, y^*_i|\hat{P}^*)$, $i = 1, \ldots, n$, where the reference set is S^*_n. [6.2] Reflect the n estimates $D(x^*_i, y^*_i|\hat{P}^*)$ about unity. [6.3] Use the computational shortcut in (10.35) to draw n bootstrap values D^{**}_i, $i = 1, \ldots, n$, from the kernel estimate of the density of the $D(x^*_i, y^*_i|\hat{P}^*)$. [6.4] Construct a pseudodataset S^{**}_n with elements (x^{**}_i, y^{**}_i) given by $y^{**}_i = D^{**}_i y^*_i / D(x^*_i, y^*_i|\hat{P}^*)$ and $x^{**}_i = x^*_i$. [6.5] Use (10.25) (or an analog of (10.25) if the convex cone estimators were used in step [1]) to compute the bootstrap estimate $D(x_0, y_0|\hat{P}^{**})$; here the reference set S^{**}_n. [6.6] Repeat steps [6.3]–[6.5] B_2 times to obtain a set of B_2 bootstrap values

$$\left\{D_{b_2}(x_0, y_0\Big|\hat{P}^{**})\right\}_{b_2=1}^{B_2}.$$

[6.7] Use (10.43) to determine $\hat{a}^*_\alpha$ and $\hat{b}^*_\alpha$, and then use these in (10.44) together with the estimate $D(x_0, y_0|\hat{P}^*)$ obtained in step [6] to obtain I^*_α, an estimated confidence interval for $D(x_0, y_0|\hat{P})$.

[7] Repeat steps [4]–[6.7] B_1 times to obtain a set of B_1 bootstrap estimates

$$\left\{D_{b_1}(x_0, y_0|\hat{P}^*)\right\}_{b_1=1}^{B_1},$$

and a set of confidence intervals $\left\{I^*_{\alpha,b_1}\right\}_{b_1=1}^{B_1}$.

[8] Compute the proportion of cases where I^*_{∂,b_1} covers $D(x_0, y_0|\hat{P})$:

$$\hat{\pi}B_1B_2(\alpha) = \frac{1}{B_1}\sum I(D(x_0, y_0|\hat{P}) \in I^*_{\alpha,b_1}),$$

where $I(\cdot)$ is the indicator function.

[9] Solve the equation $\hat{\pi}B_1B_2(\alpha) = 1 - \alpha_0$ for α, where $1 - \alpha_0$ is the desired, nominal coverage. Denote the solution by $\hat{\alpha}$. [10] Use (10.28) to determine $\hat{a}_{\hat{\alpha}}$ and $\hat{b}_{\hat{\alpha}}$, and then use these in (10.30) together with the original estimate $D \times (x_0, y_0|\hat{P})$ obtained in step [2] to obtain $I_{\hat{\alpha}}$, the estimated confidence interval for $D(x_0, y_0|P)$. In addition, the bootstrap estimates can be used with (10.32) to obtain an estimate of the bias of $D(x_0, y_0|\hat{P})$, and with (10.33) to obtain a bias-corrected estimator if the condition in (6.13) is satisfied.

10.10 Conclusions

As noted in Sect. 10.1, DEA methods are still quite young compared to many existing parametric statistical techniques. Because of their nonparametric nature and the resulting lack of structure, obtaining asymptotic results is difficult, and even when they can be obtained, their practical use remains to be demonstrated. The bootstrap is a natural tool to apply in such cases. Our Monte Carlo evidence confirms, however, that the structure of the underlying, true model plays a crucial role in determining how well the bootstrap will perform in a given applied setting. In those cases, where the researcher does not have access to the true model as we did in our Monte Carlo experiments, the iterated bootstrap offers a convenient, analogous approach for evaluating the performance of the bootstrap and providing corrections if needed.

The bootstrap methods described in this chapter are widely used. By now, the statistical properties of DEA estimators are well understood, even in the full multivariate case with $p > 1$ and $q > 1$, due to the results in Kneip et al. (2008), and the need for inference due to uncertainty surrounding estimation is also well-understood. The FEAR software library (Wilson 2008) allows empirical researchers to apply these methods with minimal effort. With the continuing decline in the cost of computing power, it is reasonable to expect that these methods will become a standard part of DEA applications. In fact, a search on August 23, 2010 using

Google Scholar with the keywords "DEA," "bootstrap," and "efficiency" returned 2,620 references, while a search using "FDH" in place of "DEA" returned 526 references.

Acknowledgments Léopold Simar gratefully acknowledges the Research support from "Projet d'Actions de Recherche Concertées" (No. 98/03-217) and from the "Inter-university Attraction Pole," Phase V (No. P5/24) from the Belgian Government.

Paul W. Wilson also gratefully acknowledges Research support from the Texas Advanced Computing Center at the University of Texas, Austin.

References

Banker RD. Maximum likelihood, consistency and data envelopment analysis: a statistical foundation. Manag Sci. 1993;39(10):1265–73.

Banker RD, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manag Sci. 1984;30:1078–92.

Beran R, Ducharme G. Asymptotic theory for bootstrap methods in statistics. Montreal, QC: Centre de Reserches Mathematiques, University of Montreal; 1991.

Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. Ann Stat. 1981;9:1196–217.

Chambers RG, Chung Y, Färe R. Benefit and distance functions. Econ Theor. 1996;70:407–19.

Charnes A, Cooper WW, Rhodes E. Measuring the inefficiency of decision making units. Eur J Oper Res. 1978;2(6):429–44.

Deprins D, Simar L, Tulkens H. Measuring labor inefficiency in post offices. In: Marchand M, Pestieau P, Tulkens H, editors. The performance of public enterprises: concepts and measurements. Amsterdam: North-Holland; 1984. p. 243–67.

Efron B. Bootstrap methods: another look at the jackknife. Ann Stat. 1979;7:1–16.

Efron B. The Jackknife, the Bootstrap and other Resampling plans, CBMS-NSF regional conference series in applied mathematics, #38. Philadelphia: SIAM; 1982.

Efron B, Tibshirani RJ. An introduction to the Bootstrap. London: Chapman and Hall; 1993.

Färe R. Fundamentals of production theory. Berlin: Springer; 1988.

Färe R, Grosskopf S, Lovell CAK. The measurement of efficiency of production. Boston: Kluwer-Nijhoff; 1985.

Farrell MJ. The measurement of productive efficiency. J Roy Stat Soc A. 1957;120:253–81.

Gijbels I, Mammen E, Park BU, Simar L. On estimation of monotone and concave frontier functions. J Am Stat Assoc. 1999;94:220–8.

Hall P. The Bootstrap and Edgeworth expansion. New York, NY: Springer; 1992.

Hall P, Härdle W, Simar L. Iterated bootstrap with application to frontier models. J Product Anal. 1995;6:63–76.

Jeong SO, Simar L. Linearly interpolated FDH efficiency score for nonconvex frontiers. J Multivar Anal. 2006;97:2141–61.

Kneip A, Park BU, Simar L. A note on the convergence of nonparametric DEA estimators for production efficiency scores. Economet Theor. 1998;14:783–93.

Kneip A, Simar L, Wilson PW. Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models. Economet Theor. 2008;24:1663–97.

Kneip A, Simar L, Wilson PW. A computationally efficient, consistent bootstrap for inference with non-parametric DEA estimators, Computational Economics (2011, in press).

Korostelev A, Simar L, Tsybakov AB. Efficient estimation of monotone boundaries. Ann Stat. 1995a;23:476–89.

Korostelev A, Simar L, Tsybakov AB. On estimation of monotone and convex boundaries, Publications de l'Institut de Statistique des Universités de Paris XXXIX 1. 1995b. 3–18.

Lewis PA, Goodman AS, Miller JM. A pseudo-random number generator for the System/360. IBM Syst J. 1969;8:136–46.
Löthgren M. How to Bootstrap DEA Estimators: A Monte Carlo Comparison (contributed paper presented at the Third Biennal Georgia Productivity Workshop, University of Georgia, Athens, GA, October 1998), Working paper series in economics and Finance #223, Department of Economic Statistics, Stockhold School of Economics, Sweden. 1998.
Löthgren M. Bootstrapping the Malmquist productivity index-A simulation study. Appl Econ Lett. 1999;6:707–10.
Löthgren M, Tambour M. Bootstrapping the DEA-based Malmquist productivity index, in essays on performance measurement in health care, Ph.D. dissertation by Magnus Tambour. Stockholm, Sweden: Stockholm School of Economics; 1997.
Löthgren M, Tambour M. Testing scale efficiency in DEA models: a bootstrapping approach. Appl Econ. 1999;31:1231–7.
Manski CF. Analog estimation methods in econometrics. New York: Chapman and Hall; 1988.
Park B, Simar L, Weiner C. The FDH estimator for productivity efficiency scores: asymptotic properties. Economet Theor. 1999;16:855–77.
Park BU, Jeong S-O, Simar L. Asymptotic distribution of conical-hull estimators of directional edges. Ann Stat. 2010;38:1320–40.
Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical recipes. Cambridge: Cambridge University Press; 1986.
Scott DW. Multivariate density estimation. New York, NY: Wiley; 1992.
Shephard RW. Theory of cost and production function. Princeton: Princeton University Press; 1970.
Silverman BW. Choosing the window width when estimating a density. Biometrika. 1978;65:1–11.
Silverman BW. Density estimation for statistics and data analysis. London: Chapman & Hall Ltd.; 1986.
Simar L, Wilson PW. Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. Manag Sci. 1998;44(11):49–61.
Simar L, Wilson PW. Some problems with the Ferrier/Hirschberg bootstrap idea. J Product Anal. 1999a;11:67–80.
Simar L, Wilson PW. Of course we can bootstrap DEA scores! But does it mean anything? Logic trumps wishful thinking. J Product Anal. 1999b;11:93–7.
Simar L, Wilson PW. Estimating and bootstrapping Malmquist indices. Eur J Oper Res. 1999c;115:459–71.
Simar L, Wilson PW. Statistical inference in nonparametric frontier models: the state of the art. J Product Anal. 2000a;13:49–78.
Simar L, Wilson PW. A general methodology for bootstrapping in nonparametric frontier models. J Appl Stat. 2000b;27:779–802.
Simar L, Wilson PW. Nonparametric tests of returns to scale. Eur J Oper Res. 2002;139:115–32.
Simar L, Wilson PW. Statistical inference in nonparametric frontier models: Recent developments and perspectives. In: Fried H, Lovell CAK, Schmidt S, editors. The measurement of productive efficiency, chapter 4. 2nd ed. Oxford: Oxford University Press; 2008. p. 421–521.
Simar L, Wilson PW. 2009 Inference by subsampling in nonparametric frontier models. Discussion paper #0933, Institut de Statistique, Universit´e Catholique de Louvain, Louvain-la-Neuve, Belgium.
Spanos A. Statistical foundations of econometric modelling. Cambridge: Cambridge University Press; 1986.
Swanepoel JWH. A note on proving that the (modified) bootstrap works. Comm Stat Theor Meth. 1986;15:3193–203.
Wheelock DC, Wilson PW. Explaining bank failures: deposit insurance, regulation, and efficiency. Rev Econ Stat. 1995;77:689–700.

Wheelock DC, Wilson PW. Why do banks disappear? The determinants of US bank failures and acquisitions. Rev Econ Stat. 2000;82:127–38.

Wilson PW. Testing independence in models of productive efficiency. J Prod Anal. 2003;20:361–90.

Wilson PW. FEAR: a software package for frontier efficiency analysis with R. Soc Econ Plann Sci. 2008;42:247–54.

Wilson PW. Asymptotic properties of some non-parametric hyperbolic efficiency estimators. In: van Keilegom I, Wilson PW, editors. Exploring research frontiers in contemporary statistics and econometrics. Berlin: Physica-Verlag; 2010.

Chapter 11
Statistical Tests Based on DEA Efficiency Scores

Rajiv D. Banker and Ram Natarajan

Abstract This chapter is written for analysts and researchers who may use data envelopment analysis (DEA) to statistically evaluate hypotheses about characteristics of production correspondences and factors affecting productivity. Contrary to some characterizations, it is shown that DEA is a full-fledged statistical methodology, based on the characterization of DMU efficiency as a stochastic variable. The DEA estimator of the production frontier has desirable statistical properties, and provides a basis for the construction of a wide range of formal statistical tests (Banker RD Mgmt Sci. 1993;39(10):1265–73). Specific tests described here address issues such as comparisons of efficiency of groups of DMUs, existence of scale economies, existence of allocative inefficiency, separability and substitutability of inputs in production systems, analysis of technical change and productivity change, impact of contextual variables on productivity, and the adequacy of parametric functional forms in estimating monotone and concave production functions.

Keywords Data envelopment analysis • Statistical tests

11.1 Introduction

Data envelopment analysis (DEA) continues to be used extensively in many settings to analyze factors influencing the efficiency of organizations. The DEA approach specifies the production set only in terms of desirable properties such as convexity and monotonicity, without imposing any parametric structure on it (Banker et al. 1984). Despite its widespread use, many persons continue to classify DEA inappropriately as a nonstatistical approach. However, many recent

R. Natarajan (✉)
School of Management, The University of Texas at Dallas, Richardson, TX 75083–0688, USA
e-mail: nataraj@utdallas.edu

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*, International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_11,

advances have established the statistical properties of the DEA efficiency estimators. Based on statistical representations of DEA, rigorous statistical tests of various hypotheses have also been developed.

To start with, Banker (1993) provided a formal statistical foundation for DEA by identifying conditions under which DEA estimators are statistically consistent and maximize likelihood. He also developed hypothesis tests for efficiency comparison when a group of DMUs (decision-making units) is compared with another. Since the publication of Banker (1993), a number of significant advances have been made in developing DEA-based hypothesis tests to address a wide spectrum of issues of relevance to users of DEA. These include issues such as efficiency comparison of groups, existence of scale inefficiency, impact of contextual variables on productivity, adequacy of parametric functional forms in estimating monotone and concave production functions, examination of input separability and substitutability in production systems, existence of allocative inefficiency, and evaluation of technical change and productivity change. In the rest of this chapter, we describe different DEA-based tests of hypotheses in the form of a reference list for researchers and analysts interested in applying DEA to efficiency measurement and production frontier estimation.[1]

The chapter proceeds as follows. In the next section, Sect. 11.2, we present statistical tests that are relevant for applications of DEA in environments where the deviation from the frontier is caused by a single one-sided stochastic random variable representing DMU inefficiency. We describe salient aspects of the statistical foundation provided in Banker (1993), discuss hypothesis tests for efficiency comparison of groups of DMUs, for the existence of scale inefficiency or allocative inefficiency and for input substitutability. In Sect. 11.3, we address situations where the production frontier shifts over time. We describe DEA-based techniques and statistical tests to evaluate and test for existence of productivity and technical change over time. In Sect. 11.4, we discuss the application of DEA in environments where the deviation from the production frontier arises as a result of two stochastic variables, one representing inefficiency and the other random noise. We explain how efficiency comparisons across groups of DMUs can be carried out, and describe DEA-based methods to determine the impact of contextual or environmental variables on inefficiency. We also suggest methods to evaluate the adequacy of an assumed parametric form for situations where prior guidance specifies only monotonicity and concavity for a functional relationship. Finally, we summarize and conclude in Sect. 11.5.

[1]Bootstrapping, discussed in Chap. 10, offers an alternative approach for statistical inference within nonparametric efficiency estimation. Extensive Monte Carlo evidence, however, indicates that the Bootstrap-based tests do not perform as well as the tests described here and the performance of the Bootstrap-based tests does not justify the considerably greater computational burden of conducting those tests.

11.2 Hypothesis Tests When Inefficiency is the Only Stochastic Variable

The tests described in this section build on the work in Banker (1993) that provides a formal statistical basis for DEA estimation techniques. We briefly describe the salient aspects of Banker (1993) before discussing the hypothesis tests.

11.2.1 Statistical Foundation for DEA

Consider observations on $j = 1, \ldots, N$ DMUs, each observation comprising a vector of outputs $\mathbf{y}_j \equiv (y_{1j}, \ldots, y_{Rj}) \geq 0$ and a vector of inputs $\mathbf{x}_j \equiv (x_{1j}, \ldots, x_{Ij}) \geq 0$, where $y \in Y$ and $\mathbf{x} \in X$, and Y and X are convex subsets of $\Re^R$ and $\Re^I$, respectively. Input quantities and output mix proportion variables are random variables.[2] The production correspondence between the frontier output vector $\mathbf{y}^0$ and the input vector $\mathbf{x}^0$ is represented by the production frontier $g(\mathbf{y}^0, \mathbf{x}^0) = 0$. The support set of the frontier is monotonically increasing and convex set $T \equiv \{(\mathbf{y}, \mathbf{x}) | \mathbf{y}$ can be produced from $\mathbf{x}\}$. The inefficiency of a specific DMU_j is, $\theta_j \equiv \max\{\theta | (\theta \mathbf{y}_j, \mathbf{x}_j) \in T\}$. It is modeled as a scalar random variable that takes values in the range $[1, \infty)$ and is distributed with the probability density function $f(\theta)$ in this range. Banker (1993) imposes additional structure on the distribution of the inefficiency variable requiring that there is a nonzero likelihood of nearly efficient performance, i.e., $\int_1^{1+\delta} f(\theta) \mathrm{d}\theta > 0$ for all $\delta > 0$.

In empirical applications of DEA, the inefficiency variable θ is not observed and needs to be estimated from output and input data. The following Banker, Charnes, and Cooper (BCC 1984) linear program is used to estimate the inefficiency:

$$\hat{\theta}_j = \text{argmax} \left\{ \theta \left| \begin{array}{l} \sum_{k=1}^{N} \lambda_k y_{rk} \geq \theta y_{rj}, \quad \forall r = 1, \ldots, R; \quad \sum_{k=1}^{N} \lambda_k x_{ik} \leq x_{ij}, \quad \forall i = 1, \ldots, I; \\ \sum_{k=1}^{N} \lambda_k = 1, \quad \lambda_k \geq 0, \quad \forall k = 1, \ldots, N \end{array} \right. \right\} \tag{11.1}$$

where $\hat{\theta}_j$ is the DEA estimator of θ_j.

Modeling the inefficiency deviation as a stochastic variable that is distributed independently of the inputs, enabled Banker (1993) to derive several results that

[2] Alternative specifications that specify output quantities and input mix proportion as random variables or endogenous input mix decisions based on input prices modeled as random variables can also be used.

provide a statistical foundation for hypothesis tests using DEA. He demonstrated that the DEA estimators of the true inefficiency values maximize likelihood provided the density function of the inefficiency random variable $f(\theta)$ is monotone decreasing. He also pointed out that a broad class of probability distributions, including the exponential and half-normal distributions, possesses monotone decreasing density functions. Banker (1993) also shows that the DEA estimator of the inefficiency underestimates the true inefficiency in finite samples. More importantly, he shows that asymptotically this bias reduces to zero; that is the DEA estimators are consistent if the probability of observing nearly efficient DMUs is strictly positive.[3]

While consistency is a desirable property of an estimator, it does not by itself guide the construction of hypothesis tests. However, Banker (1993) exploits the consistency result to prove that for "large" samples the DEA estimators of inefficiency for any given subset of DMUs follow the same probability distribution as the true inefficiency random variable. This is, perhaps, the most important result in the Banker (1993) paper since it implies that, for large samples, distributional assumptions imposed for the true inefficiency variable can be carried over to the empirical distribution of the DEA estimator of inefficiency and test statistics based on the DEA estimators of inefficiency can be evaluated against the assumed distribution of the true inefficiency.

11.2.2 Efficiency Comparison of Two Groups of DMUs

The asymptotic properties of the DEA inefficiency estimator are used by Banker (1993) to construct statistical tests enabling a comparison of two groups of DMUs to assess whether one group is more efficient than the other. Banker (1993) proposes parametric as well as nonparametric tests to evaluate the null hypothesis of no difference in the inefficiency distributions of two subsamples, G_1 and G_2, that are part of the sample of N DMUs when the sample size, N, is large. For N_1 and N_2 DMUs in subgroups G_1 and G_2, respectively, the null hypothesis of no difference in inefficiency between the two subgroups can be tested using the following procedures:

1. If the logarithm[4] of the true inefficiency θ_j is distributed as exponential over $[0, \infty)$ for the two subgroups, then under the null hypothesis that there is no

[3] The consistency result does not require that the probability density function $f(\theta)$ be monotone decreasing. It only requires that there is a positive probability of observing nearly efficient DMUs, which is a much weaker condition than the monotonicity condition required for the DEA estimators to be maximum likelihood.

[4] Alternatively, the assumption may be maintained that $t(\theta_j)$ is distributed as exponential over $[0, \infty)$ where $t(\cdot)$ is some specified transformation function. Then the test statistic is given by $\left[\sum_{j \in G_1} t(\hat{\theta}_j)/N_1\right] \Big/ \left[\sum_{j \in G_2} t(\hat{\theta}_j)/N_2\right]$.

difference between the two groups, the test statistic is calculated as $\left[\sum_{j\in G_1} \ln(\hat{\theta}_j)/N_1\right] \Big/ \left[\sum_{j\in G_2} \ln(\hat{\theta}_j)/N_2\right]$ and evaluated relative to the critical value of the F distribution with $(2N_1, 2N_2)$ degrees of freedom.

2. If logarithm of the true inefficiency θ_j is distributed as half-normal over the range $[0, \infty)$ for the two subgroups, then under the null hypothesis that there is no difference between the two groups, the test statistic is calculated as $\left[\sum_{j\in G_1} \left\{\ln(\hat{\theta}_j)\right\}^2/N_1\right] \Big/ \left[\sum_{j\in G_2} \left\{\ln(\hat{\theta}_j)\right\}^2/N_2\right]$ and evaluated relative to the critical value of the F distribution with (N_1, N_2) degrees of freedom.
3. If no such assumptions are maintained about the probability distribution of inefficiency, a nonparametric Kolmogorov–Smirnov's test statistic given by the maximum vertical distance between $F^{G_1}(\ln(\hat{\theta}_j))$ and $F^{G_2} \ln(\hat{\theta}_j)$, the empirical distributions of $\ln(\hat{\theta}_j)$ for the groups G_1 and G_2, respectively, is used. This statistic, by construction, takes values between 0 and 1 and a high value for this statistic is indicative of significant differences in inefficiency between the two groups.

11.2.3 Tests of Returns to Scale

Examining the existence of increasing or decreasing returns to scale is an issue of interest in many DEA studies. We provide a number of DEA-based tests to evaluate returns to scale using DEA inefficiency scores.[5] Consider the inefficiency $\hat{\theta}_j^C$ estimated using the CCR (Charnes et al. 1978) model obtained from the BCC linear program in (11.1) by deleting the constraint $\sum_{k=1}^{N} \lambda_k = 1$, i.e.,

$$\hat{\theta}_j^C = \text{argmax}\left\{\theta \,\middle|\, \begin{array}{l} \sum_{k=1}^{N} \lambda_k y_{rk} \geq \theta y_{rj}, \quad \forall r = 1,\ldots,R; \quad \sum_{k=1}^{N} \lambda_k x_{ik} \leq x_{ij}, \quad \forall i = 1,\ldots,I; \\ \lambda_k \geq 0, \quad \forall k = 1,\ldots,N \end{array}\right\}. \tag{11.2}$$

[5] Chapter 2 of this handbook provides a detailed discussion of qualitative and quantitative aspects of returns to scale in DEA. Banker et al. (2010a) present Monte Carlo evidence revealing that these tests outperform Bootstrap-based tests with respect to both Type 1 and Type 2 errors. The performance of all such tests declines considerably if the number of sample observations is small relative to the dimensionality of the production set measured as the number of inputs plus the number of outputs.

By construction, $\hat{\theta}_j^C \geq \hat{\theta}_j$. Scale inefficiency is then estimated as $\hat{\theta}_j^S = \hat{\theta}_j^C / \hat{\theta}_j$. Values of scale inefficiency significantly greater than 1 indicate the presence of scale inefficiency to the extent operations deviate from the *most productive scale size (MPSS)* (Banker 1984; Banker et al. 1984). All observations in the sample are scale efficient if and only if the sample data can be rationalized by a production set exhibiting constant returns to scale.

Under the null hypothesis of no scale inefficiency (or equivalently, under the null hypothesis of constant returns to scale), $\hat{\theta}_j^C$ is also a consistent estimator of θ_j (Banker 1993, 1996). The null hypothesis of no scale inefficiency in the sample can be evaluated by constructing the following test statistics:

1. If the logarithm of the true inefficiency θ_j is distributed as exponential over $[0, \infty)$, then under the null hypothesis of constant returns to scale, the test statistic is calculated as $\sum \ln(\hat{\theta}_j^C) / \sum \ln(\hat{\theta}_j)$. This test statistic is evaluated relative to the half-F distribution $|F_{2N,2N}|$ with $2N$, $2N$ degrees of freedom over the range $[1, \infty)$, since by construction the test statistic is never less than 1. The half-F distribution is the F distribution truncated below at 1, the median of the F distribution when the two degrees of freedom are equal.
2. If the logarithm of the true inefficiency θ_j is distributed as half-normal over the range $[0, \infty)$, then under the null hypothesis of constant returns to scale, the test statistic is calculated as $\sum \left\{ \ln(\hat{\theta}_j^C) \right\}^2 \Big/ \sum \left\{ \ln(\hat{\theta}_j) \right\}^2$ and evaluated relative to the half-F distribution $|F_{N,N}|$ with N, N degrees of freedom over the range $[1, \infty)$.
3. If no such assumptions are maintained about the probability distribution of inefficiency, a nonparametric Kolmogorov–Smirnov's test statistic given by the maximum vertical distance between $F^C(\ln(\hat{\theta}_j^C))$ and $F(\ln(\hat{\theta}_j))$, the empirical distributions of $\ln(\hat{\theta}_j^C)$ and $\ln(\hat{\theta}_j)$, respectively, is used. This statistic, by construction, takes values between 0 and 1 and a high value for this statistic is indicative of the existence of significant scale inefficiency in the sample.

The above tests evaluate the null hypothesis of constant returns to scale against the alternative of variable returns to scale. In addition, it is also possible to test the null hypothesis of nondecreasing returns to scale against the alternative of decreasing returns to scale and the null hypothesis of nonincreasing returns to scale against the alternative of increasing returns to scale. Two additional inefficiency estimators $\hat{\theta}_j^D$ and $\hat{\theta}_j^E$ required for these tests are calculated by solving the program in (11.1) after changing the constraint $\sum_{k=1}^N \lambda_k = 1$ to $\sum_k \lambda_k \leq 1$ for $\hat{\theta}_j^D$ and to $\sum_k \lambda_k \geq 1$ for $\hat{\theta}_j^E$. By construction, $\hat{\theta}_j^C \geq \hat{\theta}_j^D \geq \hat{\theta}_j$ and $\hat{\theta}_j^C \geq \hat{\theta}_j^E \geq \hat{\theta}_j$.

The following is a test of the null hypothesis of nondecreasing returns to scale against the alternative of decreasing returns to scale:

1. If the logarithm of the true inefficiency θ_j is distributed as exponential over $[0, \infty)$, the test statistic is calculated as $\sum \ln(\hat{\theta}_j^E) / \sum \ln(\hat{\theta}_j)$ or $\sum \ln(\hat{\theta}_j^C) / \sum \ln(\hat{\theta}_j^D)$. Each of these statistics is evaluated relative to the half-F distribution $|F_{2N,2N}|$ with $2N$, $2N$ degrees of freedom over the range $[1, \infty)$.

2. If the logarithm of the true inefficiency θ_j is distributed as half-normal over the range $[0, \infty)$, the test statistic is calculated as either $\sum \left\{\ln(\hat{\theta}_j^E)\right\}^2 \Big/ \sum \left\{\ln(\hat{\theta}_j)\right\}^2$ or $\sum \left\{\ln(\hat{\theta}_j^C)\right\}^2 \Big/ \sum \left\{\ln(\hat{\theta}_j^D)\right\}^2$ and evaluated relative to the half-F distribution $|F_{N,N}|$ with N, N degrees of freedom over the range $[1, \infty)$.
3. If no such assumptions are maintained about the probability distribution of inefficiency, a nonparametric Kolmogorov–Smirnov's test statistic given by either the maximum vertical distance between F^E $(\ln(\hat{\theta}_j^E))$ and F $(\ln(\hat{\theta}_j))$, or that between F^C $(\ln(\hat{\theta}_j^C))$ and F^D $(\ln(\hat{\theta}_j^D))$ is used.

The test statistics for testing the null of nonincreasing returns to scale against the alternative of increasing returns to scale can be developed in a similar fashion by interchanging $\hat{\theta}_j^E$ and $\hat{\theta}_j^D$ in the statistics above.

11.2.4 Tests of Allocative Efficiency

In this section, we describe DEA-based tests that can be used to examine the existence of allocative inefficiencies associated with input utilization. In many DEA studies that have examined inefficiency associated with input utilization, inefficiency is often estimated using aggregate cost expenditure information. Banker et al. (2004) address the situation when information about input prices is not available, except for the knowledge that the firms procure the inputs in the same competitive market place. They employ the result that the DEA technical inefficiency measure using a single aggregate cost variable, constructed from multiple inputs weighted by their unit prices, reflects the aggregate technical and allocative inefficiency. This result is then used to develop statistical tests of the null hypothesis of no allocative inefficiency analogous to those of the null hypothesis of no scale inefficiency described earlier.

For the purposes of this section, consider observations on $j = 1, \ldots, N$ DMUs, each observation comprising an output vector $\mathbf{y}_j \equiv (y_{1j}, \ldots, y_{Rj}) \geq 0$ and a vector of input costs $\mathbf{c}_j \equiv (c_{1j}, \ldots, c_{Ij}) \geq 0$ for $i = 1, \ldots, I$ inputs. Each input i, $i = 1, \ldots, I$, is bought by all firms in the same competitive market at a price p_i. Let $\mathbf{p} = (p_1, \ldots, p_I)$ be the vector of input prices. The cost of input i for DMU_j is then $c_{ij} = p_i x_{ij}$. The total cost of inputs for DMU_j is $c_j = \sum p_i x_{ij} = \sum c_{ij}$. The input quantities x_{ij} and the price vector $\mathbf{p}$ are not observable by the researcher.[6] Only the output and cost information are observed.

[6] Sections 1.5 and 1.6 discuss efficiency estimation for situations in which unit prices and unit costs are available.

The aggregate technical and allocative inefficiency estimator, $\hat{\theta}_j^Z \geq 1$, is estimated using the following linear program that utilizes output and aggregate cost data:

$$\hat{\theta}_j^Z = \text{argmax}\left\{ \theta \;\middle|\; \begin{array}{l} \sum_{k=1}^{N} \lambda_k y_{rk} \geq y_{rj}, \quad \forall r = 1, \ldots, R; \qquad \sum_{k=1}^{N} \lambda_k c_k \leq c_j/\theta; \\ \sum_{k=1}^{N} \lambda_k = 1\,, \quad \lambda_k \geq 0, \quad \forall k = 1, \ldots, N \end{array} \right\}. \tag{11.3}$$

The technical inefficiency estimator $\hat{\theta}_j^B \geq 1$ is estimated as

$$\hat{\theta}_j^B = \text{argmax}\left\{ \theta \;\middle|\; \begin{array}{l} \sum_{k=1}^{N} \lambda_k y_{rk} \geq y_{rj}, \quad \forall r = 1, \ldots, R; \quad \sum_{k=1}^{N} \lambda_k c_{ik} \leq c_{ij}/\theta, \quad \forall i = 1, \ldots, I; \\ \sum_{k=1}^{N} \lambda_k = 1\,, \quad \lambda_k \geq 0, \quad \forall k = 1, \ldots, N \end{array} \right\}, \tag{11.4}$$

where $\hat{\theta}_j^B$ is a consistent estimator of the true technical inefficiency θ_j^B (Banker 1993). Further the estimator for the allocative inefficiency, $\hat{\theta}_j^V$, can be calculated as $\hat{\theta}_j^Z / \hat{\theta}_j^B$. Under the null hypothesis that the sample data does not exhibit any allocative inefficiency, $\hat{\theta}_j^Z$ is also a consistent estimator of the true technical inefficiency θ_j^B. This leads to the following tests of the null hypothesis of no allocative inefficiency in the utilization of the inputs as opposed to the alternative of existence of allocative inefficiency:

1. If $\ln(\theta_j^B)$ is distributed as exponential over $[0, \infty)$, then under the null hypothesis of no allocative inefficiency, the test statistic is calculated as $\sum_{j=1}^{N} \ln(\hat{\theta}_j^Z) \Big/ \sum_{j=1}^{N} \ln(\hat{\theta}_j^B)$ and evaluated relative to the critical value of the half-F distribution with $(2N, 2N)$ degrees of freedom.
2. If $\ln(\theta_j^B)$ is distributed as half-normal over the range $[0, \infty)$, then under the null hypothesis of no allocative inefficiency, the test statistic is calculated as $\sum_{j=1}^{N} (\ln(\hat{\theta}_j^Z))^2 \Big/ \sum_{j=1}^{N} (\ln(\hat{\theta}_j^B))^2$ and evaluated relative to the critical value of the half-F distribution with (N, N) degrees of freedom.
3. If no such assumptions are maintained about the probability distribution of inefficiency, a nonparametric Kolmogorov–Smirnov's test statistic given by the maximum vertical distance between $F(\ln(\hat{\theta}_j^Z))$ and $F(\ln(\hat{\theta}_j^B))$, the empirical distributions of $\ln(\hat{\theta}_j^Z)$ and $\ln(\hat{\theta}_j^B)$, respectively, is used. This statistic, by

construction, takes values between 0 and 1 and a high value is indicative of the existence of allocative inefficiency.

There could also be situations involving multiple outputs and multiple inputs where output quantity information may not be available but monetary value of the individual outputs along with input quantity information may be available. In such situations, output-based allocative inefficiency can be estimated and tested using procedures similar to those outlined above for input-based allocative efficiency (Banker et al. 2007).

11.2.5 Tests of Input Separability

In this section, we describe DEA-based tests that can be used to evaluate the null hypothesis of input separability, i.e., the influence of each of the inputs on the output is independent of other inputs, against the alternative hypothesis that the inputs are substitutable. The DEA-based tests proposed in this section evaluate the null hypothesis of input separability over the entire sample data in contrast to the parametric (Berndt and Wood 1975) tests which are operationalized only at the sample mean.

Once again, consider observations on $j = 1, \ldots, N$ DMUs, each observation comprising an output vector $\mathbf{y}_j \equiv (y_{1j}, \ldots, y_{Rj}) \geq 0$, a vector of inputs $\mathbf{x}_j \equiv (x_{1j}, \ldots, x_{Ij}) \geq 0$ and a production technology characterized by a monotone increasing and convex production possibility set $T \equiv \{(\mathbf{y}, \mathbf{x}) | \mathbf{y}$ can be produced from $\mathbf{x}\}$. The input-oriented inefficiency measure for this technology is estimated using the following BCC – Banker et al. (1984) – linear program:

$$\hat{\theta}_j^{\text{SUB}} = \text{argmax} \left\{ \theta \;\middle|\; \begin{array}{l} \sum_{k=1}^{N} \lambda_k y_{rk} \geq y_{rj}, \quad \forall r = 1, \ldots, R; \quad \sum_{k=1}^{N} \lambda_k x_{ik} \leq x_{ij}/\theta, \quad \forall i = 1, \ldots, I; \\ \sum_{k=1}^{N} \lambda_k = 1, \quad \lambda_k \geq 0, \quad \forall k = 1, \ldots, N \end{array} \right\}. \tag{11.5}$$

When the inputs are separable, input inefficiency is first estimated considering only one input at a time, resulting in I different inefficiency measures corresponding to the I inputs. The overall DMU inefficiency is then estimated as the minimum of these I inefficiency measures. Specifically, the inefficiency corresponding to input i is measured as

$$\hat{\theta}_j^i = \text{argmax}\left\{ \theta_i \;\middle|\; \begin{array}{l} \sum_{k=1}^{N} \lambda_k y_{rk} \geq y_{rj}, \quad \forall r = 1,\ldots,R; \qquad \sum_{k=1}^{N} \lambda_k x_{ik} \leq x_{ij}/\theta_i\,; \\ \sum_{k=1}^{N} \lambda_k = 1\,, \quad \lambda_k \geq 0, \quad \forall k = 1,\ldots,N \end{array} \right\}. \tag{11.6}$$

The inefficiency measure under the input separability assumption is then estimated as $\hat{\theta}_j^{\text{SEP}} = \text{Min}\left\{\hat{\theta}_j^i |\; i = 1,\ldots,I\right\}$. Since $\hat{\theta}_j^{\text{SEP}}$ is estimated from a less constrained program, $\hat{\theta}_j^{\text{SEP}} \geq \hat{\theta}_j^{\text{SUB}}$. Under the null hypothesis of input separability, the asymptotic empirical distributions of $\hat{\theta}_j^{\text{SEP}}$ and $\hat{\theta}_j^{\text{SUB}}$ are identical, with each retrieving the distribution of the true input inefficiency θ.

The above discussion leads to the following tests of the null hypothesis of separability in the utilization of the inputs as opposed to the alternative of substitutability of inputs:

1. If $\ln(\theta_j)$ is distributed as exponential over $[0, \infty)$, then under the null hypothesis of separability of inputs, the test statistic is calculated as $\sum_{j=1}^{N} \ln(\hat{\theta}_j^{\text{SEP}}) \Big/ \sum_{j=1}^{N} \ln(\hat{\theta}_j^{\text{SUB}})$ and evaluated relative to the critical value of the half-F distribution with $(2N, 2N)$ degrees of freedom.
2. If $\ln(\theta_j)$ is distributed as half-normal over the range $[0, \infty)$, then under the null hypothesis of input separability, the test statistic is calculated as $\sum_{j=1}^{N} \left(\ln(\hat{\theta}_j^{\text{SEP}})\right)^2 \Big/ \sum_{j=1}^{N} \left(\ln(\hat{\theta}_j^{\text{SUB}})\right)^2$ and evaluated relative to the critical value of the half-F distribution with (N, N) degrees of freedom.
3. If no such assumptions are maintained about the probability distribution of inefficiency, a nonparametric Kolmogorov–Smirnov's test statistic given by the maximum vertical distance between $F(\ln(\hat{\theta}_j^{\text{SEP}}))$ and $F(\ln(\hat{\theta}_j^{\text{SUB}}))$, the empirical distributions of $\ln(\hat{\theta}_j^{\text{SEP}})$ and $\ln(\hat{\theta}_j^{\text{SUB}})$, respectively, is used. This statistic, by construction, takes values between 0 and 1 and a high value is indicative of the rejection of input separability in the production technology.

11.3 Hypothesis Tests for Situations Characterized by Shifts in Frontier

In the previous section, we presented DEA tests that are useful in situations where cross-sectional data on DMUs is used for efficiency analysis. In this section, we describe estimation procedures and statistical tests when both longitudinal and

cross-sectional data on DMUs are available and where the object of interest is change in productivity over time.[7] Productivity researchers have advocated both parametric and nonparametric approaches to estimate and analyze the impact of technical change and efficiency change on productivity change. The nonparametric literature (e.g., Färe et al. 1997; Ray and Desli 1997; Førsund and Kittelsen 1998) has focused exclusively on the measurement of productivity change using DEA without attempting to provide a statistical basis to justify those methods. Most empirical applications of these methods have sought to test whether significant technical change occurred, on average, for the sample observations, or whether the technical change for a subsample was significantly different from that for another subsample. Banker et al. (2009) have developed DEA-based estimation methods and tests of productivity change and technical change that justify some common methods employed in prior empirical research. This section summarizes salient aspects of the Banker et al. (2009) study. For ease of exposition, we focus on a single output rather than a multiple output vector to illustrate the techniques and tests in this section.

Let $\mathbf{x_t} = (x_{1t}, \ldots, x_{it}, \ldots, x_{It}) \in X$, $x_{it} > 0$, $t = 0, 1$ be the I-dimensional input vector in period t. Consider two possible values of the time subscript t corresponding to a base period ($t = 0$) and another period ($t = 1$). The production correspondence in time t between the frontier output y_t^* and the input $\mathbf{x_t}$, is represented as

$$y_t^* = \phi^t(\mathbf{x_t}), \quad t = 0, 1. \tag{11.7}$$

The function $\phi^t(\cdot)$: $X \rightarrow \Re^+$ is constrained to be monotone increasing and concave in the input $\mathbf{x_t}$ but not required to follow any specific parametric functional form. The input vector $\mathbf{x_t}$, the production function $\phi^t(\mathbf{x_t})$, and a relative efficiency random variable, α_t, that takes values between $-\infty$ and 0, together determine the realized output, y_t, in period t.[8] Specifically,

$$y_t \equiv e^{\alpha_t}\phi^t(\mathbf{x_t}) \tag{11.8}$$

For the above setup, estimators of technical change, relative efficiency change, and productivity change, respectively, $\hat{b}_j^{(N)}$, $\hat{r}_j^{(N)}$, and $\hat{g}_j^{(N)}$, for the jth DMU can be estimated from observed output values and *estimators* of frontier outputs using the following expressions:

[7] The treatment of productivity change in this section provides an alternative to treatments using the Malmquist index described in Chap. 8 and, additionally, has the advantage of providing explicit statistical characterizations.

[8] The relative efficiency random variable α_t and the inefficiency measure θ_t are linked by the relationship $\alpha_t = -\ln(\theta_t)$.

$$\hat{b}_j^{(N)} = \{\ln(\hat{\phi}^1(x_{j1})) - \ln(\hat{\phi}^0(x_{j1}))\}$$
$$\hat{r}_j^{(N)} = \{\ln(y_{j1}/\hat{\phi}^1(x_{j1})) - \ln(y_{j0}/\hat{\phi}^0(x_{j0}))\}$$
$$\hat{g}_j^{(N)} = \{\ln(y_{j1}/\hat{\phi}^0(x_{j1})) - \ln(y_{j0}/\hat{\phi}^0(x_{j0}))\}. \tag{11.9}$$

These estimators satisfy the fundamental relationship that productivity change is the sum of technical change and relative efficiency change, i.e., $\hat{g}_j^{(N)} = \hat{b}_j^{(N)} + \hat{r}_j^{(N)}$. The frontier outputs required for the estimation of the various change measures in (11.9) are estimated using linear programs. The estimation of $\hat{\phi}^0(x_{j0})$ and $\hat{\phi}^1(x_{j1})$, is done using only the input–output observations from the base period or period 1, as the case may be. The linear program for estimating $\hat{\phi}^0(x_{j0})$ is the following Banker et al. (BCC 1984) model:

$$\hat{\phi}^0(x_{j0}) = \text{argmax}\left\{\hat{\phi} \;\middle|\; \begin{array}{ll} \sum_{k=1}^{N} \lambda_{k0}^0 y_{k0} \geq \hat{\phi}; & \sum_{k=1}^{N} \lambda_{k0}^0 x_{ik0} \leq x_{ij0}, \quad \forall i = 1,\ldots,I; \\ \sum_{k=1}^{N} \lambda_{k0}^0 = 1\,, & \lambda_{k0}^0 \geq 0, \quad \forall k = 1,\ldots,N \end{array}\right\}. \tag{11.10}$$

A similar program is used to estimate $\hat{\phi}^1(x_{j1})$ from input and output data from period 1. $\hat{\phi}^0(x_{j0})$ and $\hat{\phi}^1(x_{j1})$ are consistent estimators of $\phi^0(\mathbf{x}_{j0})$ and $\phi^1(\mathbf{x}_{j1})$, respectively, (Banker 1993).

The base period frontier value, $\phi^0(\mathbf{x}_{j1})$, corresponding to *period 1* input, $\mathbf{x}_{j1}$, is estimated based on the following linear program:

$$\hat{\phi}^0(x_{j1}) = \text{argmax}\left\{\hat{\phi} \;\middle|\; \begin{array}{ll} \sum_{k=1}^{N} \lambda_{k0}^0 y_{k0} \geq \hat{\phi}; & \sum_{k=1}^{N} \lambda_{k0}^0 x_{ik0} \leq x_{ij1}, \quad \forall i = 1,\ldots,I; \\ \sum_{k=1}^{N} \lambda_{k0}^0 = 1\,, & \lambda_{k0}^0 \geq 0, \quad \forall k = 1,\ldots,N \end{array}\right\} \tag{11.11}$$

or $= y_{j1}$ when the above linear program is not feasible

Note that the difference between the above model and the traditional BCC model is that the observation under evaluation is not included in the reference set for the constraints in (11.11) as in the super-efficiency model described first in Banker et al. (1989) and Anderson and Petersen (1993). It is the case that $\hat{\phi}^0(x_{j1})$ is a consistent estimator of $\phi^0(\mathbf{x}_{j1})$.

Given the N values of technical, relative efficiency and productivity changes estimated using (11.9), the estimators of the medians of these performance measures are

$$\hat{b}^{\mathrm{MD}(N)} = \operatorname{argmin} \frac{1}{N} \sum_{j=1}^{N} |\hat{b}_j^{(N)} - b|$$

$$\hat{r}^{\mathrm{MD}(N)} = \operatorname{argmin} \frac{1}{N} \sum_{j=1}^{N} |\hat{r}_j^{(N)} - r|$$

$$\hat{g}^{\mathrm{MD}(N)} = \operatorname{argmin} \frac{1}{N} \sum_{j=1}^{N} |\hat{g}_j^{(N)} - g|. \tag{11.12}$$

Banker et al. (2009) show that $\hat{b}^{\mathrm{MD}(N)}$, $\hat{r}_j^{\mathrm{MD}(N)}$, and $\hat{g}_j^{\mathrm{MD}(N)}$ are consistent estimators of the population median technical change β^{MD}, population median relative efficiency change ρ^{MD}, and population median productivity change γ^{MD}, respectively.

Consider the number of observations, $\hat{p}_{\beta}^{(N)}$, $\hat{p}_{\rho}^{(N)}$, and $\hat{p}_{\gamma}^{(N)}$ (out of the sample of N observations) for which $\hat{b}_j^{(N)}$, $\hat{r}_j^{(N)}$, and $\hat{g}_j^{(N)}$, respectively, is strictly positive. Tests for the median of the various performance measures being equal to zero are conducted as follows:

1. Under the null hypothesis of zero median technical change between the base period and period t, i.e., $\beta^{\mathrm{MD}} = 0$, the statistic $\hat{p}_{\beta}^{(N)}$ is asymptotically distributed as a binomial variate with parameters N and 0.5, i.e., $\hat{p}_{\beta}^{(N)} \sim b(N, 0.5)$.
2. Under the null hypothesis of zero median relative efficiency change between the base period and period t, i.e., $\rho^{\mathrm{MD}} = 0$, the statistic $\hat{p}_{\rho}^{(N)}$ is asymptotically distributed as a binomial variate with parameters N and 0.5, i.e., $\hat{p}_{\rho}^{(N)} \sim b(N, 0.5)$.
3. Under the null hypothesis of zero median technical change between the base period and period t, i.e., $\gamma^{\mathrm{MD}} = 0$, the statistic $\hat{p}_{\gamma}^{(N)}$ is asymptotically distributed as a binomial variate with parameters N and 0.5, i.e., $\hat{p}_{\gamma}^{(N)} \sim b(N, 0.5)$.

Banker et al. (2009) also provide methods for estimating and testing the location of the population mean of the various performance measures. The mean technical change $\widehat{\overline{b}}^{(N)}$, mean relative efficiency change $\widehat{\overline{r}}^{(N)}$, and mean productivity change $\hat{\overline{g}}^{(N)}$ are estimated as

$$\widehat{\overline{b}}^{(N)} = \frac{1}{N} \sum_{j=1}^{N} \left(\ln(\hat{\phi}^1(x_{j1})) - \ln(\hat{\phi}^0(x_{j1})) \right)$$

$$\hat{\overline{r}}^{(N)} = \frac{1}{N} \left\{ \ln\left(y_{j1} / \hat{\phi}^1(x_{j1}) \right) - \ln\left(y_{j0} / \hat{\phi}^0(x_{j0}) \right) \right\}$$

$$\hat{\overline{g}}^{(N)} = \frac{1}{N} \left\{ \ln\left(y_{j1} / \hat{\phi}^0(x_{j1}) \right) - \ln\left(y_{j0} / \hat{\phi}^0(x_{j0}) \right) \right\}, \tag{11.13}$$

where $\widehat{\overline{b}}^{(N)}$, $\hat{\overline{r}}^{(N)}$, and $\hat{\overline{g}}^{(N)}$ are consistent estimators of the population mean technical change, $\overline{\beta}$, population mean relative efficiency change, $\overline{\rho}$, and population productivity change, $\overline{\gamma}$, respectively.

Next consider the statistics $\hat{t}^{(N)}(\beta) = \sqrt{N}\widehat{\overline{b}}^{(N)}/\hat{s}(\beta)$, $\hat{t}^{(N)}(\rho) = \sqrt{N}\widehat{\overline{r}}^{(N)}/\hat{s}(\rho)$, and $\hat{t}^{(N)}(\gamma) = \sqrt{N}\hat{\overline{g}}^{(N)}/\hat{s}(\gamma)$ where

$$\hat{s}^2(\beta) = \left(\sum_{j=1}^{N}\left(\ln(\hat{\phi}^1(x_{j1})) - \ln(\hat{\phi}^0(x_{j1}))\right)^2 - N\left(\widehat{\overline{b}}^{(N)}\right)^2\right)/(N-1)$$

$$\hat{s}^2(\rho) = \left(\sum_{j=1}^{N}\left(\ln(y_{j1}/\hat{\phi}^1(x_{j1})) - \ln(y_{j0}/\hat{\phi}^0(x_{j0}))\right)^2 - N\left(\widehat{\overline{r}}^{(N)}\right)^2\right)/(N-1)$$

$$\hat{s}^2(\gamma) = \left(\sum_{j=1}^{N}\left(\ln\left(y_{j1}/\hat{\phi}^0(x_{j1})\right) - \ln(y_{j0}/\hat{\phi}^0(x_{j0}))\right)^2 - N\left(\hat{\overline{g}}^{(N)}\right)^2\right)/(N-1). \tag{11.14}$$

For large samples, the distribution of each of $\hat{t}^{(N)}(\beta)$, $\hat{t}^{(N)}(\rho)$, and $\hat{t}^{(N)}(\gamma)$ approaches that of a Student's T variate with $N-1$ degrees of freedom. Therefore, a simple t-test for the mean of the DMU-specific estimators for the various performance measures estimated using (11.9) is appropriate when the sample size is large.

Banker et al. (2005) apply the estimation procedures and tests described above to examine components of productivity change in the public accounting industry toward the end of the twentieth century.[9] Banker et al. (2009) present extensive Monte Carlo simulation evidence that indicates that the performance of test statistics derived from these computationally simple, one-shot productivity and technical change estimators is robust to the presence of noise, small sample sizes, as well as when the production correspondence involves multiple inputs.[10]

11.4 Hypothesis Tests for Composed Error Situations

The tests described in the previous sections are conditioned on a data-generating process (DGP) that characterizes the deviation of the actual output from the production frontier as arising only from a stochastic inefficiency term. In this section, we describe the application of DEA-based tests for composed error situations where the DGP involves not only the one-sided inefficiency term but also a noise term that is independent of the inefficiency. Recently, Gstach (1998) and Banker and Natarajan (2008) have developed DEA-based estimation procedures for environments

[9] Zhang et al. (2011) apply the procedures to analyze patterns of change in the institutes of the Chinese Academy of Sciences following its introduction of a new research policy matched by extensive improvements in organizational processes.

[10] Banker et al. (2009) also present Monte Carlo evidence indicating that the Bootstrap-based method to construct confidence intervals for each observation is inappropriate for testing average technical change over all observations.

characterized by both inefficiency and noise. The tests developed by Banker (1993) can be adapted to these environments through an appropriate transformation of the inefficiency term. We describe these tests below.

11.4.1 Tests for Efficiency Comparison

Consider observations on $j = 1,\ldots, N$ DMUs, each observation comprising a single output $y_j \geq 0$ and a vector of inputs $\mathbf{x}_j \equiv (x_{1j},\ldots,x_{Ij}) \geq 0$. The production correspondence between the frontier output y^0 and the I inputs is represented as $y^0 = g(\mathbf{x})$ subject to the assumption that $g(\cdot)$ is monotonically increasing and concave in $\mathbf{x}$. The deviation from the frontier for the jth DMU could be positive or negative and is represented as $\varepsilon_j = u_j - v_j = g(\mathbf{x}_j) - y_j$. Thus, the deviation is modeled as the sum of two components, a one-sided inefficiency term, u_j, and a two-sided random noise term v_j bounded above at V^{M}, analogous to composed error formulations in parametric stochastic frontier models (Aigner et al. 1977; Meeusen and van den Broeck 1977; Banker and Natarajan 2008). In this stochastic framework, Banker et al. (2010b) propose five statistical tests to compare the efficiency of two groups of DMUs.

As before, consider two subsamples, G_1 and G_2, that are part of the sample of N DMUs when the sample size, N, is large. Let the true inefficiency, u_j, be distributed with means $\bar{u}_1$ and $\bar{u}_2$ in the two groups. Further assume that the variance of the inefficiency is the same in both groups. Define $\tilde{u}_j = V^{\mathrm{M}} - v_j + u_j$. We can estimate $\hat{\tilde{u}}_j$ by applying DEA on input and output data from the full sample of N DMUs to obtain a consistent estimator of $\tilde{u}_j$. For N_1 and N_2 DMUs in subgroups G_1 and G_2, respectively, the null hypothesis of no difference in mean inefficiency between the two subgroups can be tested using the following procedures:

1. Consider the OLS regression $\hat{\tilde{u}}_j = a_0 + a_1 z_j + e_j$ estimated using a total of $N_1 + N_2$ DEA inefficiency scores. z_j is a dummy variable that takes a value of 0 if a particular DMU belongs to group G_1 and 1 if it belongs to G_2 and e_j is an i.i.d. error term. The regression coefficient $\hat{a}_1$ is a consistent estimator of $\bar{u}_2 - \bar{u}_1$, the difference in mean inefficiency between groups G_2 and G_1. The t-statistic associated with this regression coefficient can be used to evaluate whether the two groups are significantly different in terms of mean inefficiency.
2. Assume that the probability distributions of the inefficiency random variable u_j and the noise random variable v_j are such that $\tilde{u}_j = V^{\mathrm{M}} - v_j + u_j$ is distributed as a log-normal variable in the two groups. Under the null hypothesis that the mean inefficiencies are equal, i.e., $\bar{u}_1 = \bar{u}_2$ and assuming that the variance of u_j is the same in the two groups, the Student-t statistic $\hat{t} = (\bar{\tilde{I}}_1 - \bar{\tilde{I}}_2)\Big/\hat{S}\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}$ distributed with $(N_1 + N_2 - 2)$ degrees of freedom can be used to evaluate the null hypothesis of no difference in mean inefficiency across the two groups.

$\bar{\hat{I}}_1$ is $\frac{1}{N_1}\sum_{j=1}^{N_1}\ln(\hat{\tilde{u}}_{j1})$, $\bar{\hat{I}}_2 = \frac{1}{N_2}\sum_{j=1}^{N_2}\ln(\hat{\tilde{u}}_{j2})$, and $\hat{S} = \left(\frac{1}{N_1+N_2-2}\left\{\sum_{j=1}^{N_1}\left\{\ln(\hat{\tilde{u}}_{j1}) - \bar{\hat{I}}_1\right\}^2 + \sum_{j=1}^{N_2}\left\{\ln(\hat{\tilde{u}}_{j2}) - \bar{\hat{I}}_2\right\}^2\right\}\right)^{0.5}$.

Banker et al. (2010) also suggest three nonparametric tests for efficiency comparison including a median test, the Mann–Whitney test, and the Kolmogorov–Smirnov test. All three tests are based on order statistics. The first of these test statistics for examining the equality of the median inefficiencies of the two groups is $\hat{Z} = (\hat{P}_1 - \hat{P}_2)\Big/\sqrt{\hat{P}(1-\hat{P})(1/N_1 + 1/N_2)}$ where $\hat{P} = (N_1\hat{P}_1 + N_2\hat{P}_2)/(N_1 + N_2)$, $\hat{P}_1 = n_1/N_1$, $\hat{P}_2 = n_2/N_2$ and n_1, n_2, respectively, are the number of observations in groups 1 and 2 that have values less than the sample median inefficiency $\hat{M}$. The statistic $\hat{Z}$is asymptotically distributed as a standard-Normal variate.

The second order statistic is the Mann–Whitney statistic which is based on the number of times $\hat{\tilde{u}}_i$ precedes $\hat{\tilde{u}}_j$ in the combined and ordered sample of the two groups $i = 1, \ldots, N_1$ and $j = 1, \ldots, N_2$. Define a random variable $\hat{D}_{ij} = \left\{1 \text{ if } \hat{\tilde{u}}_i < \hat{\tilde{u}}_j \text{ and 0 otherwise}\right\}$. Then the statistic is $\hat{U} = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\hat{D}_{ij}$ and $\hat{Z} = (\hat{U} - N_1N_2/2)\Big/\sqrt{N_1N_2(N+1)/12}$ asymptotically follows a standard-Normal distribution. Finally, the Kolmogorov–Smirnov's test statistic, calculated as the maximum vertical distance between $F^1(\hat{\tilde{u}}_1)$ and $F^2(\hat{\tilde{u}}_2)$, the empirical distributions of $\tilde{u}$ for groups 1 and 2 respectively, can also be used to make an inefficiency comparison across the two groups.

Banker et al. (2010a, b) provide extensive Monte Carlo simulation results that confirm that DEA estimators can help detect efficiency differences across groups of DMUs. The simulation experiments suggest that the five tests described above perform better than the *F*-tests in Banker (1993), described in Sect. 11.2.2, when noise plays a significant role in the DGP. On the other hand, the Banker (1993) tests are effective when efficiency dominates noise. The simulation results also suggest that the test statistics described above have the desired asymptotic properties in that their empirical distributions approach the theoretical ones.

11.4.2 Tests for Evaluating the Impact of Contextual Variables on Efficiency

Analysis of factors contributing to efficiency differences has been an important area of research in DEA. Ray (1991), for instance, regresses DEA scores on a variety of socioeconomic factors to identify key performance drivers in school districts. The two-stage approach of first calculating productivity scores and then seeking to correlate these scores with various explanatory variables has been in use for over 20 years but explanations of productivity differences using DEA are still dominated

by ad hoc speculations (Førsund 1999). Banker and Natarajan (2008, 2009) provide a general framework for the evaluation of contextual variables affecting productivity by considering a variety of DGPs and present appropriate estimation methods and statistical tests under each DGP. In this section, we describe the DEA-based tests developed in Banker and Natarajan (2008, 2009) that can be used to determine the impact of contextual or environmental variables on efficiency.

Consider observations on $j = 1, \ldots, N$ DMUs, each observation comprising a single output $y_j \geq 0$, a vector of inputs $\mathbf{x}_j \equiv (x_{1j}, \ldots, x_{Ij}) \geq 0$, and a vector of contextual variables $\mathbf{z}_j \equiv (z_{1j}, \ldots, z_{Sj})$ that may influence the overall efficiency in transforming the inputs into the outputs. The production function $g(\cdot)$ is monotone increasing and concave in $\mathbf{x}$, and relates the inputs and contextual variables to the output as specified by the equation:

$$y_j = g(\mathbf{x}_j) + v_j - h(\mathbf{z}_j) - u_j, \tag{11.15}$$

where v_j is a two-sided random noise term bounded above at V^{M}, $h(\mathbf{z}_j)$ is a nonnegative monotone increasing function, convex in $\mathbf{z}$, and u_j is a one-sided inefficiency term. The inputs, contextual variables, noise, and inefficiency are all distributed independent of each other. Defining $\tilde{g}(x_j) = g(\mathbf{x}_j) + V^{\mathrm{M}}$ and $\tilde{\delta}_j = (V^{\mathrm{M}} - v_j) + h(\mathbf{z}_j) + u_j \geq 0$, (11.15) can be expressed as

$$y_j = \tilde{g}(x_j) - \tilde{\delta}_j. \tag{11.16}$$

Since $\tilde{g}(\cdot)$ is derived from $g(\cdot)$ by multiplication with a positive constant, $\tilde{g}(\cdot)$ is also monotone increasing and concave. Therefore, the DEA inefficiency estimator, $\hat{\tilde{\delta}}_j$, obtained by performing DEA on the output–inputs observations $(y_j, \mathbf{x}_j)$, $j = 1, \ldots, N$, is a consistent estimator of $\tilde{\delta}_j$ (Banker 1993). This consistency result is used by Banker and Natarajan (2008, 2009) to develop the following DEA-based tests corresponding to different specifications of $h(\cdot)$, the function linking contextual variables to inefficiency.

Consider the case where $h(\mathbf{z}) = h(\mathbf{z};\boldsymbol{\beta})$ and $h(\mathbf{z};\boldsymbol{\beta})$ is a nonnegative function, monotone increasing in $\mathbf{z}$, linear in $\boldsymbol{\beta}$. In this case, the impact of contextual variables can be consistently estimated by regressing the first stage DEA estimate $\hat{\tilde{\delta}}_j$ on the various contextual variables associated with the various components of the $\boldsymbol{\beta}$ vector. This procedure yields consistent estimators of the parameter vector $\boldsymbol{\beta}$. In the special case where $h(\mathbf{z};\beta) = \mathbf{z}'\beta = \sum_{j=1}^{N} z_j\beta_j$, the independent variables in the regression are the same as the S contextual variables.

Banker and Natarajan (2008) identify conditions under which (a) a two-stage procedure consisting of DEA followed by ordinary least squares (OLS) regression analysis and (b) DEA in the first stage followed by maximum likelihood estimation (MLE) in the second stage yield consistent estimators of the impact of contextual variables. They point out that a necessary requirement is that the contextual variables should be independent of the input variables, but the contextual variables

may be correlated with each other. Monte Carlo simulation results in Banker and Natarajan (2008) indicate that DEA-based procedures with OLS, maximum likelihood, or even Tobit estimation in the second stage perform as well as the best of the parametric methods in the estimation of the impact of contextual variables on productivity.

Banker and Natarajan (2009) examine situations where no additional structure can be placed on $h(\mathbf{z}_j)$ except that it is a nonnegative monotone increasing function, convex in $\mathbf{z}$. Let $\tilde{\varepsilon}_j = (V^M - v_j) + u_j$. Then $\tilde{\delta}_j = h(\mathbf{z}_j) + \tilde{\varepsilon}_j$ and a second stage DEA estimation on the pseudo "input–outputs" observations $(\tilde{\delta}_j, z_j)$ yields a consistent estimator $\hat{\tilde{\varepsilon}}_j$ for $\tilde{\varepsilon}_j$ (Banker 1993). This estimator is obtained by solving the following linear programming formulation analogous to the BCC model (Banker et al. 1984) in DEA individually for each observation in the sample:

$$\hat{\psi}_j = \text{argmin}\left\{ \psi \;\middle|\; \begin{array}{ll} \sum_{k=1}^{N} \lambda_k \hat{\tilde{\delta}}_k = \psi; & \sum_{k=1}^{N} \lambda_k z_{ik} \geq z_{ij}, \quad \forall i = 1, \ldots, S; \\ \sum_{k=1}^{N} \lambda_k = 1, & \lambda_k \geq 0, \quad \forall k = 1, \ldots, N \end{array} \right\}. \tag{11.17}$$

A consistent estimator for $\tilde{\varepsilon}_j$ for each observation in the sample is obtained as $\hat{\tilde{\varepsilon}}_j = \hat{\tilde{\delta}}_j - \hat{\psi}_j$.

To evaluate the statistical significance of individual z_s, a third stage DEA estimation is first performed on the pseudo-observations $(\hat{\tilde{\delta}}_j, z_j^{-S})$ where $\mathbf{z}^{-s}$ is the original $\mathbf{z}$ vector without the z_s variable. The following modified version of the program in (11.17) is used for this purpose:

$$\hat{\psi}_j^{-s} = \text{argmin}\left\{ \psi \;\middle|\; \begin{array}{ll} \sum_{k=1}^{N} \lambda_k \hat{\tilde{\delta}}_k = \psi; & \sum_{k=1}^{N} \lambda_k z_{ik} \geq z_{ij}, \quad \forall i = 1, \ldots, s-1, s+1, \ldots, S; \\ \sum_{k=1}^{N} \lambda_k = 1, & \lambda_k \geq 0, \quad \forall k = 1, \ldots, N \end{array} \right\}. \tag{11.18}$$

Let the resulting estimator of $\tilde{\varepsilon}_j$ be $\hat{\tilde{\varepsilon}}_j^{-s} = \hat{\tilde{\delta}}_j - \hat{\psi}_j^{-s}$. Since (11.18) is a less constrained program than (11.17), $\hat{\tilde{\varepsilon}}_j^{-s} \geq \hat{\tilde{\varepsilon}}_j$ for all observations $j = 1, \ldots, N$.

Under the null hypothesis that the marginal impact of z_s (i.e., $\partial h(\mathbf{z})/\partial z_s$ if $h(\cdot)$ is differentiable) is zero, the asymptotic distributions of $\tilde{\varepsilon}$ and $\tilde{\varepsilon}^{-s}$ are identical (Banker 1993). If the asymptotic distribution of $\tilde{\varepsilon}_j$ is assumed to be exponential or half-normal, the null hypothesis of no impact of z_s is tested by comparing the ratios $\sum_{j=1}^{N} \hat{\tilde{\varepsilon}}_j^{-s} \Big/ \sum_{j=1}^{N} \hat{\tilde{\varepsilon}}_j$ or $\sum_{j=1}^{N} [\hat{\tilde{\varepsilon}}_j^{-s}]^2 \Big/ \sum_{j=1}^{N} [\hat{\tilde{\varepsilon}}_j]^2$ against critical values obtained from half-F distributions with $(2N, 2N)$ or (N, N) degrees of freedom, respectively. If $\tilde{\varepsilon}_j$ is exponentially distributed, the test statistic is evaluated relative to the half-F

distribution$|F_{2N,2N}|$ with $2N$, $2N$ degrees of freedom over the range $[1, \infty)$, since by construction the test statistic is never less than 1. If $\tilde{\varepsilon}_j$ is a half-Normal variate then the test statistic is evaluated relative to the half-F distribution $|F_{N,N}|$ with N, N degrees of freedom over the range $[1, \infty)$. Recall that $\tilde{\varepsilon} = u + (V^{\mathrm{M}} - v)$. Therefore, statistics based on Banker (1993) may not be valid unless the variance of the noise term v is considerably smaller than the variance of the inefficiency term, u, and u is distributed with a mode at its lower support.

The Kolmogorov–Smirnov statistic, which is based on the maximum vertical distance between the empirical cumulative distribution of $\hat{\tilde{\varepsilon}}_j^{-s}$ and $\hat{\tilde{\varepsilon}}_j$, can also be used to check whether the empirical distributions of $\hat{\tilde{\varepsilon}}_j^{-s}$ and $\hat{\tilde{\varepsilon}}_j$ are significantly different. If they are significantly different, then it can be established that z_s has a significant impact on productivity.

11.4.3 Tests for Evaluating the Adequacy of Parametric Functional Forms

While DEA provides a theoretically correct way to estimate monotone and concave (or convex) functional relationships, it is often useful to represent the relationship in a more parsimonious functional form that is afforded by a parametric specification. Specific parametric functional forms, such as the Cobb–Douglas, are useful if they provide a good approximation to the general monotone and concave (or convex) function as evidenced by sample data. In this section, we present methods developed in Banker et al. (2002) to evaluate the adequacy of a parametric functional form to represent the functional relationship between an endogenous variable and a set of exogenous variables given the minimal maintained assumption of monotonicity and concavity.

Consider sample data on an endogenous variable and I exogenous variables for N observations. For the jth observation, denote the endogenous variable as y_j and the vector of exogenous variables as $X_j \equiv (x_{1j}, x_{2j}, \ldots, x_{Ij})$. The relationship between y_j and X_j is specified as

$$y_j = g(X_j)\mathrm{e}^{\varepsilon_j} \tag{11.19}$$

where $g(\cdot)$ is a monotone increasing and concave function. It is assumed further that ε_j is independent of X_j and i.i.d. with a probability density function $f(\varepsilon)$ over the range $[-S_{\mathrm{L}}, S_{\mathrm{U}}] \subseteq \Re$, where $S_{\mathrm{L}} \geq 0$, and $S_{\mathrm{U}} \geq 0$ are unknown parameters that describe the lower and upper supports of the distribution. Define $\tilde{g}(X) = g(X)\mathrm{e}^{S_{\mathrm{U}}}$ and $\tilde{\varepsilon}_j = S_{\mathrm{U}} - \varepsilon_j \geq 0$ such that (11.19) can be rewritten as $y_j = \tilde{g}(X_j)e^{-\tilde{\varepsilon}_j}$. The DEA estimator of $\tilde{g}(X_j)$ can be estimated using the following linear program:

$$\hat{\tilde{g}}^{\mathrm{DEA}}(X_j) = \operatorname{argmax}\left\{ y \;\middle|\; \begin{array}{ll} \displaystyle\sum_{k=1}^{N} \lambda_k y_k = y; & \displaystyle\sum_{k=1}^{N} \lambda_k x_{ik} \le x_{ij}, \quad \forall i = 1, \ldots, I; \\ \displaystyle\sum_{k=1}^{N} \lambda_k = 1\,, & \lambda_k \ge 0, \quad \forall k = 1, \ldots, N \end{array} \right\}. \tag{11.20}$$

An estimator for $\tilde{\varepsilon}_j$ for each observation in the sample can then be obtained as $\hat{\tilde{\varepsilon}}_j^{\mathrm{DEA}} = \ln(\hat{\tilde{g}}^{\mathrm{DEA}}(X_j)) - \ln(y_j)$. Further, $\hat{\tilde{g}}^{\mathrm{DEA}}(X_j)$ and $\hat{\tilde{\varepsilon}}_j^{\mathrm{DEA}}$ are consistent estimators of $\tilde{g}(X_j)$ and $\tilde{\varepsilon}_j$, respectively (Banker 1993).

The parametric estimation is carried out by specifying a parametric form $g(X;\beta)$ and regressing $\ln(y)$ on the exogenous variables in $\ln(g(X;\beta))$. The residuals from the regression are used to obtain $\hat{S}_{\mathrm{U}} = \max\{\ln(y) - \ln(\hat{g}(X_j;\hat{\beta}))\}$ which is a consistent estimator of S_{U} (Greene 1980). The estimated deviation from the parametric frontier is then calculated as $\hat{\tilde{\varepsilon}}_j^{\mathrm{PARAM}} = \ln(\hat{g}(X_j;\hat{\beta})) + \hat{S}_{\mathrm{U}} - \ln(y_j)$. In addition to $\hat{\tilde{\varepsilon}}^{\mathrm{DEA}}$, $\hat{\tilde{\varepsilon}}_j^{\mathrm{PARAM}}$ is also a consistent estimator of $\tilde{\varepsilon}_j$ under the null hypothesis that $g(X;\beta) = g(X)$ for all X. Banker et al. (2002) prove that the asymptotic distribution of $\hat{\tilde{\varepsilon}}_j^{\mathrm{PARAM}}$ retrieves the true distribution of $\tilde{\varepsilon}$ if the parametric specification is, in fact, the true specification of the production function (i.e., $g(X;\beta) = g(X)$ for all X). Further, they also show that if the parametric specification is, in fact, the true specification of the production function then as $N \to \infty$, (a) the asymptotic distribution of $\hat{\tilde{\varepsilon}}_j^{\mathrm{PARAM}}$ converges to that of $\hat{\tilde{\varepsilon}}^{\mathrm{DEA}}$ and (b) both $\hat{\tilde{\varepsilon}}_j^{\mathrm{PARAM}}$ and $\hat{\tilde{\varepsilon}}_j^{\mathrm{DEA}}$ converge asymptotically to $\tilde{\varepsilon}_j$ for all $j \in J$, where J is a given set of observations. Based on these results, they suggest the following four tests for testing the adequacy of the parametric functional form:

1. The first test uses the Kolmogorov–Smirnov test statistic given by the maximum vertical distance between $\hat{F}(\hat{\tilde{\varepsilon}}_j^{\mathrm{DEA}})$ and $\hat{F}(\hat{\tilde{\varepsilon}}_j^{\mathrm{PARAM}})$, where $\hat{F}(\hat{\tilde{\varepsilon}}_j^{\mathrm{DEA}})$ and $\hat{F}(\hat{\tilde{\varepsilon}}_j^{\mathrm{PARAM}})$ denote the empirical distributions of $\hat{\tilde{\varepsilon}}_j^{\mathrm{DEA}}$ and $\hat{\tilde{\varepsilon}}_j^{\mathrm{PARAM}}$ respectively. A low value is indicative of support for the null hypothesis that $g(X;\beta)$ adequately represents $g(X)$.
2. The second procedure is based on the regression of rank of $\hat{\tilde{\varepsilon}}_j^{\mathrm{DEA}}$ on the rank of $\hat{\tilde{\varepsilon}}_j^{\mathrm{PARAM}}$ (Iman and Conover 1979). Under the null hypothesis that the parametric form is adequate, the expected value of the coefficient on $\hat{\tilde{\varepsilon}}_j^{\mathrm{PARAM}}$ in the rank regression is asymptotically equal to 1. The null hypothesis is evaluated against the alternative hypothesis that the regression coefficient has a value less than 1.
3. The third test procedure employs the Wilcoxon rank-sum test to evaluate whether the empirical distributions $\hat{F}(\hat{\tilde{\varepsilon}}_j^{\mathrm{DEA}})$ and $\hat{F}(\hat{\tilde{\varepsilon}}_j^{\mathrm{PARAM}})$ are different. If the test shows these distributions to be different, the adequacy of the parametric form is rejected.
4. The fourth procedure is based on Theil's (1950) distribution-free test. This test evaluates the null hypothesis that $\mu_1 = 1$ against the alternative $\mu_1 \neq 1$ in the relation $\hat{\tilde{\varepsilon}}^{\mathrm{DEA}} = \mu_0 + \mu_1 \hat{\tilde{\varepsilon}}^{\mathrm{PARAM}}$. To compute Theil's statistic, the difference

$D_j = \hat{\hat{\varepsilon}}_j^{\text{DEA}} - \hat{\hat{\varepsilon}}_j^{\text{PARAM}}$ is calculated and then the data are sorted by $\hat{\hat{\varepsilon}}_j^{\text{PARAM}}$. Next, a score $c_{ji} = 1, 0$, or -1 is assigned for each $i < j$ depending on whether $D_j - D_i > 0$, $=0$, or <0 respectively. Theil's test statistic C is defined as $C = \sum_{j=1}^{N} c_{ji}$. Theil's test statistic is distributed as a standard normal variate for large samples and a high absolute value of C rejects the adequacy of the parametric functional form.

Banker et al. (2002) also propose an alternative approach to evaluating the adequacy of a parametric functional form. The approach suggested by them relies on Wooldridge's (1992) Davidson–Mackinnon type test to evaluate a linear null model against a nonparametric alternative. Banker et al. (2002) propose the use of a sieve DEA estimator as the nonparametric alternative for the purposes of Wooldridge's test since the test cannot be applied directly when the nonparametric alternative is based on the traditional DEA estimator. Interested readers are referred to Sect. 11.2.3 of Banker et al. (2002) for additional details on how Wooldridge's test can be applied to examine the adequacy of a specified parametric form for a monotone and concave function.

11.5 Concluding Remarks

We have described here several statistical tests that can be used to test hypotheses of interest and relevance to applied users of DEA. A common underlying theme of these tests is that the deviation from the DEA frontier can be viewed as a stochastic variable. While the DEA estimator is biased in finite samples, the expected value of the DEA estimator is almost certainly the true parameter value in large samples. The tests described in this chapter rely on this asymptotic property of the DEA estimator.

An important caveat is that the tests described in this chapter are designed for large samples. Results of simulation studies conducted on many of the tests proposed in this study suggest that these tests perform very well for sample sizes similar to those used in many typical applications of DEA.[11] These tests need to be used with caution in small samples. We believe additional simulation studies are warranted to provide evidence on small sample performance of the tests described here. Clearly, this is an important area for future research.

We believe that there are many more avenues and areas where DEA-based statistical tests can be applied. This is because the flexible structure of DEA facilitates application in a large number of situations where insufficient information

[11] Banker (1996), Banker et al. (2010a), Banker and Natarajan (2008), Banker et al. (2009), and Banker et al. (2010b) provide details on some of these simulation studies. Based on the results of these studies, it appears that the tests described in this chapter perform well in sample sizes of the order of 50 or more, unless the dimensionality of the production set is very high.

or guidance may preclude the use of parametric methods. Statistical tests developed during the past 15–20 years have contributed significantly to the reliability of managerial and policy implications of DEA studies and we believe that they will continue to enrich future applications of DEA.

References

Aigner DJ, Lovell CAK, Schmidt P. Formulation and estimation of stochastic frontier production function models. J Econom. 1977;6:21–37.

Anderson P, Petersen NC. A procedure for ranking efficient units in data envelopment analysis. Manage Sci. 1993;39:1261–64.

Banker RD. Estimating most productive scale size using data envelopment analysis. Eur J Oper Res. 1984;17:35–44.

Banker RD. Maximum likelihood, consistency and data envelopment analysis: a statistical foundation. Mgmt Sci. 1993;39(10):1265–73.

Banker RD. Hypothesis tests using data envelopment analysis. J Prod Anal. 1996;7:139–59.

Banker RD, Charnes A, Cooper WW. Models for the estimation of technical and scale inefficiencies in data envelopment analysis. Manage Sci. 1984;30:1078–92.

Banker RD, Chang H, Natarajan R. Productivity change, technical progress and relative efficiency change in the public accounting industry. Manage Sci. 2005;51(2):291–304.

Banker RD, Chang H, Chang S. Statistical tests of returns to scale using DEA. Working paper. Temple University; 2010a.

Banker RD, Chang H, Natarajan R. Estimating technical and allocative efficiency using DEA: an application to the U.S. Public Accounting Industry. J Prod Anal. 2007;27:115–21.

Banker RD, Das S, Datar S. Analysis of cost variances for management control in hospitals. Res Governmental Nonprofit Account. 1989;5:269–91.

Banker RD, Janakiraman S, Natarajan R. Evaluating the adequacy of parametric functional forms in estimating monotone and concave production functions. J Prod Anal. 2002;17:111–32.

Banker RD, Janakiraman S, Natarajan R. Analysis of trends in technical and allocative efficiency: an application to Texas Public School Districts. Eur J Oper Res. 2004;154:477–91.

Banker RD, Natarajan R. Evaluating contextual variables affecting productivity using data envelopment analysis. Oper Res. 2008;56(1):48–58.

Banker RD, Natarajan R. DEA-based hypothesis tests to evaluate contextual variables affecting productivity. Working Paper. Temple University and the University of Texas at Dallas; 2009.

Banker RD, Natarajan R, Parthasarathy S. Nonparametric Estimation and statistical tests of components of productivity change. Working paper. Temple University and the University of Texas at Dallas; 2009.

Banker RD, Zheng Z, Natarajan R. DEA-based hypothesis tests for comparing two groups of decision making units. Eur J Oper Res. 2010b;206:231–8.

Berndt E, Wood D. Technology, prices and the derived demand for energy. Rev Econ Stat. 1975;57:259–68.

Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res. 1978;2:429–44.

Färe R, Griffel-tatje E, Grosskopf S, Lovell CAK. Biased technical change and the Malmquist productivity index. Scand J Econ. 1997;99:119–27.

Førsund FR. The evolution of DEA – the economics perspective. University of Oslo, Norway: Mimeo; 1999.

Førsund F, Kittelsen S. Productivity development of Norwegian electricity distribution utilities. Resour Energy Econ. 1998;20:207–24.

Greene WH. Maximum likelihood estimation of econometric frontier production functions. J Econom. 1980;13:27–56.
Gstach D. Another approach to data envelopment analysis in noisy environments: DEA+. J Prod Anal. 1998;9:161–76.
Iman RL, Conover WJ. The use of rank transform in regression. Technometrics. 1979;21:499–509.
Meeusen W, van den Broeck J. Efficiency estimation from Cobb–Douglas production functions with composed error. Int Econ Rev. 1977;18:435–44.
Ray S. Resource-use efficiency in public schools: a study of Connecticut data. Manage Sci. 1991;37:1620–28.
Ray S, Desli E. Productivity growth, technical progress, and efficiency change in industrialized countries: comment. Am Econ Rev. 1997;87:1033–9.
Theil H. A rank-invariant method of linear and polynomial regression analysis. I Proc Kon Ned Akad V Wetensch. 1950;A53:386–92.
Wooldridge JM. A test for functional form against nonparametric alternatives. Econom Theory. 1992;8:452–75.
Zhang D, Banker RD, Li X, Liu W. Performance impact of research policy at the Chinese Academy of Sciences. Res Policy. 2011;40(6):875–85.

Chapter 12
Modeling DMU's Internal Structures: Cooperative and Noncooperative Approaches

Wade D. Cook, Liang Liang, and Joe Zhu

Abstract An important area of development in recent years in data envelopment analysis has been the applications wherein internal structures of DMUs are considered. For example, DMUs may consist of subunits or represent two-stage processes. One particular subset of such processes is those in which all the outputs from the first stage are the only inputs to the second stage. This chapter first reviews these models and discusses relations among various approaches. Our focus here is the approaches based upon either Stackelberg (leader–follower) or cooperative game concepts. We then examine the more general problem of an open multistage process where some outputs from a given stage may leave the system while others become inputs to the next stage. As well, new inputs can enter at any stage. We then discuss the modeling of this more general network structure.

Keywords Data envelopment analysis • Efficiency • Game • Intermediate measure • Cooperative • Two-stage

12.1 Introduction

In many data envelopment analysis (DEA) settings, DMUs may consist of two-stage processes with intermediate measures. For example, Seiford and Zhu (1999) use a two-stage process to measure the profitability and marketability of US commercial banks. In their study, profitability is measured using labor and assets as inputs, and the outputs are profits and revenue. In the second stage for marketability, the profits and revenue are used as inputs, while market value, returns, and earnings per share constitute the outputs. Zhu (2000) applies the same two-stage process to the Fortune Global 500 companies. Kao and Hwang (2008) describe a two-stage process where

J. Zhu (✉)
School of Business, Worcester Polytechnic Institute, Worcester, MA 01609, USA
e-mail: jzhu@wpi.edu

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_12,

24 nonlife insurance companies use operating and insurance expenses to generate premiums in the first stage, and then underwriting and investment profits in the second stage. Other examples include the impact of information technology use on bank branches performance (Wang et al. 1997; Chen and Zhu 2004; Chen et al. 2006), two-stage Major League Baseball performance (Sexton and Lewis 2003), physician care (Chilingerian and Sherman 2004), and many others.

In all of the above examples, DMUs under evaluation share a common feature found in many two-stage processes, namely that outputs from the first stage become the inputs to the second stage. The outputs from the first stage are referred to as intermediate measures. We point out that in these examples, it is the case that the intermediate measures are the *only* inputs to the second stage, i.e., there are no additional independent inputs to that stage. There are, of course, other types of two-stage processes and even DMUs with network structures that may have inputs to the second stage in addition to the intermediate measures. For example, Liang et al. (2006) use DEA to measure the performance of supply chains with two members (as in a manufacturer–retailer setting, for example); this is also a two-stage process. See also Gaski (1984). However, the second stage (retailer) has not only the inputs from the first stage (manufacturer), but also its own inputs not linked with the first stage. Or, in a more general situation than two-stage processes, Castelli et al. (2004) discuss DMUs with two-stage and two-layer structures. The network DEA approach of Färe and Whittaker (1995) and Färe and Grosskopf (1996), and the slack-based network DEA approach of Tone and Tsutsui (2009) may involve more than two stages. The reader is also referred to Castelli et al. (2008) for a review of DEA models for cases where there are DMUs with internal structures. More recently, Chen (2009) developed a network DEA model incorporating dynamic effects in production networks.

Liang et al. (2008) use the concept of cooperative and noncooperative game theory to model DMUs that have a particular two-stage structure. (See also Kao and Hwang 2008, and Chen et al. 2009.) While these approaches can be extended to DMUs that have more than two stages, such an extension requires that the multistage processes share the unique feature that all outputs from any stage represent the only inputs to the next stage. While these *closed* systems do exist, the more prevalent case is that where each stage is *open*, that is it has its own inputs (and/or outputs) in addition to the intermediate measures.

As pointed out in Cook et al. (2010b), such open multistage structures are relatively common, particularly in processing industries. Consider, for example, the situation in which a coal mining company wishes to evaluate the efficiency of a set of collieries (mining operations) in a large coal field. Typically, the process of delivering finished products to the customer is multistage in nature. In simple terms, *Stage 1* would involve the *extraction* of the raw or run-of-mine (ROM) coal from underground or open pit coal reserves. At the mine site, the ROM is generally put through a process where screens separate the product into different size categories; e.g., a "more than one-inch diameter" category and a "less than one inch" category. The resulting "size grades," representing the outputs from this first stage, are then transported to an on-site *washing facility*, which might be deemed *Stage 2*.

The washing process filters out any material below a certain specific gravity; this portion is unsuitable for sale and is discarded. A portion of the remaining usable coal (outputs from Stage 2) is sold to the open market as a finished product, and at management's discretion (based on estimates of the demand), the remaining product is sent to *Stage 3*, the *crusher*. The crushing process also produces waste or discard, with the remaining material, sometimes referred to as "middlings," being sold or blended with other materials to make such products as briquettes. This latter process might be thought of as *Stage 4*.

Based upon additive efficiency decomposition and cooperative game theory concept, Cook et al. (2010b) develop DEA models for DMUs with general network structures.

In this chapter, we first present the DEA models for simple two-stage processes and then, generalize the results to more general network structures. See also Cook et al. (2010a) for a review on these models and their relations.

12.2 Two-Stage Processes

Consider a generic two-stage process as shown in Fig. 12.1, for each of a set of n DMUs. Using the notions in Chen and Zhu (2004) and Kao and Hwang (2008), we assume each DMU$_j$ ($j = 1, 2, \ldots, n$) has m inputs x_{ij}, ($i = 1, 2, \ldots, m$) to the first stage, and D outputs z_{dj}, ($d = 1, 2, \ldots, D$) from that stage. These D outputs then become the inputs to the second stage and will be referred to as intermediate measures. The outputs from the second stage are y_{rj}, ($r = 1, 2, \ldots, s$).

We denote the efficiency for the first stage as e_j^1 and second stage as e_j^2, for each DMU$_j$. Using the constant returns to scale (CRS) DEA model of Charnes et al. (1978), we define

$$e_j^1 = \frac{\sum_{d=1}^{D} w_d z_{dj}}{\sum_{i=1}^{m} v_i x_{ij}} \quad \text{and} \quad e_j^2 = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{d=1}^{D} \tilde{w}_d z_{dj}} \tag{12.1}$$

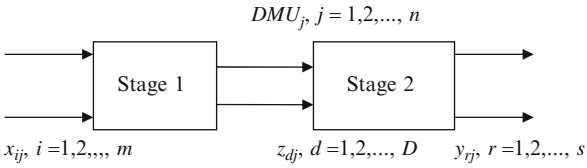


Fig. 12.1 Two-stage process

where v_i, w_d, $\tilde{w}_d$, and u_r are unknown non-negative weights. Note that w_d can be equal to $\tilde{w}_d$.

Clearly, one can apply two separate DEA runs to the two stages as in Seiford and Zhu (1999) and Zhu (2000). However, such an approach does not treat z_{dj} in a coordinated manner. For example, suppose the first stage is DEA efficient and the second stage is not. When the second stage improves its performance, by reducing the inputs z_{dj} via an input-oriented DEA model, the reduced z_{dj} may render the first stage inefficient.

It is useful to point out that given individual efficiency measures e_j^1 and e_j^2, for stages 1 and 2, respectively, it is reasonable to define the efficiency of the overall two-stage process either as $\frac{1}{2}(e_j^1 + e_j^2)$ or $e_j^1 \cdot e_j^2$. If the input-oriented DEA model is used, then we should as well require that $e_j^1 \leq 1$ and $e_j^2 \leq 1$. The above definition ensures that the two-stage process is efficient if and only if $e_j^1 = e_j^2 = 1$.

If we define $e_j = \sum_{r=1}^{s} u_r y_{ro} / \sum_{i=1}^{m} v_i x_{io}$ as the two-stage overall efficiency as in Kao and Hwang (2008), we have $e_j = e_j^1 \cdot e_j^2$ at optimality provided we assume $w_d = \tilde{w}_d$. Note that such a decomposition of efficiency is not available in the standard DEA approach of Seiford and Zhu (1999), the two-stage approach of Chen and Zhu (2004) and Chen et al. (2006), and the network DEA approach. Note also that if only one layer is considered in the internal structure of Castelli et al. (2004), then a similar efficiency decomposition can be obtained.

In Liang et al. (2006), additional inputs to the second stage are introduced. As a result, $e_j^2 = \sum_{r=1}^{s} u_r y_{rj} / \sum_{d=1}^{D} \tilde{w}_d z_{dj} + \sum_{h=1}^{H} Q_h x_{hj}^2$, where x_{hj}^2 $(h = 1, \ldots, H)$ are inputs to the second stage that are not related to the first stage. In this case, it may be more convenient to express the overall efficiency as $\frac{1}{2}(e_j^1 + e_j^2)$, since the alternative, namely $e_j^1 \cdot e_j^2$, results in a highly nonlinear problem. In Liang et al. (2006), the concepts of the Stackelberg (or leader–follower) game and the cooperative game are used to develop models for measuring the performance of supply chains. We note that their models can actually be directly applied to the two-stage process described in Fig. 12.1, since if there are no additional inputs x_{hj}^2 $(h = 1, \ldots, H)$, the structure of their two-member supply chain is identical to the two-stage process shown.

Liang et al. (2008) provide detailed models for the two-stage process using the same modeling principle as in Liang et al. (2006).

12.3 Centralized Model

Liang et al. (2006) show that using the concept of cooperative game theory or centralized control, the two stages will jointly determine a set of optimal weights on the intermediate factors to maximize their efficiency scores (as is true where the manufacturer and retailer jointly determine prices, order quantities, etc., to achieve maximum profit (Huang and Li 2001)). In other words, the cooperative or centralized approach is characterized by letting $w_d = \tilde{w}_d$ in (12.1), and the

efficiency scores of both stages are optimized simultaneously. The optimization can be based upon maximizing the average of e_o^1 and e_o^2 in a nonlinear program as in Liang et al. (2006), Kao and Hwang (2008), and Liang et al. (2008). However, it is noted that because of the assumption $w_d = \tilde{w}_d$ in (12.1), $e_o^1 \cdot e_o^2$ becomes $\sum_{r=1}^{s} u_r y_{ro} / \sum_{i=1}^{m} v_i x_{io}$. Therefore, instead of maximizing the average of e_o^1 and e_o^2, we have

$$e_o^{\text{centralized}} = \text{Max} e_o^1 \cdot e_o^2 = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}}$$
$$\text{s.t. } e_j^1 \le 1 \text{ and } e_j^2 \le 1 \text{ and } w_d = \tilde{w}_d. \tag{12.2}$$

Model (12.2) can be converted into the following linear program format:

$$e_o^{\text{centralized}} = \text{Max} \sum_{r=1}^{s} u_r y_{ro}$$
$$\text{s.t.} \sum_{r=1}^{s} u_r y_{rj} - \sum_{d=1}^{D} w_d z_{dj} \le 0, \quad j = 1,2,\ldots,n$$
$$\sum_{d=1}^{D} w_d z_{dj} - \sum_{i=1}^{m} v_i x_{ij} \le 0, \quad j = 1,2,\ldots,n$$
$$\sum_{i=1}^{m} v_i x_{io} = 1$$
$$w_d \ge 0, \quad d = 1,2,\ldots,D; \quad v_i \ge 0, i = 1,2,\ldots,m; \quad u_r \ge 0, r = 1,2,\ldots,s. \tag{12.3}$$

Model (12.3) is the Kao and Hwang (2008) model and the centralized model developed in Liang et al. (2008). Note that constraints $\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0$ are redundant in Kao and Hwang's (2008) model, since $\sum_{r=1}^{s} u_r y_{rj} - \sum_{d=1}^{D} w_d z_{dj} \le 0$ and $\sum_{d=1}^{D} w_d z_{dj} - \sum_{i=1}^{m} v_i x_{ij} \le 0$ imply $\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0$

Model (12.3) gives the overall efficiency of the two-stage process. Assume that the above model (12.3) yields a unique solution. We can then obtain

$$e_o^{1,\text{centralized}} = \frac{\sum_{d=1}^{D} w_d^* z_{do}}{\sum_{i=1}^{m} v_i^* x_{io}} = \sum_{d=1}^{D} w_d^* z_{do} \quad \text{and} \quad e_o^{2,\text{centralized}} = \frac{\sum_{r=1}^{s} u_r^* y_{ro}}{\sum_{d=1}^{D} w_d^* z_{do}} \tag{12.4}$$

as the efficiencies for the first and second stages, respectively. If we denote the optimal value to model (12.3) as $e_o^{\text{centralized}}$, then we have $e_o^{\text{centralized}} = e_o^{1,\text{centralized}} \cdot e_o^{2,\text{centralized}}$.

If only one layer is considered in the internal structure of Castelli et al. (2004), then the same above efficiency decomposition can be obtained. Therefore, the approaches of Castelli et al. (2004) and Kao and Hwang (2008) can be viewed as cooperative game models.

As noted in Kao and Hwang (2008), optimal multipliers from model (12.3) may not be unique. They propose deriving the maximum achievable value of $e_o^{1,\text{centralized}}$ or $e_o^{2,\text{centralized}}$. In fact, as shown in Liang et al. (2008), their models can also be used to test whether $e_o^{1,\text{centralized}}$ and $e_o^{2,\text{centralized}}$, obtained from model (12.3), are unique. The maximum achievable value of $e_o^{1,\text{centralized}}$ can be determined via

$$
\begin{aligned}
& e_o^{1+} = \text{Max} \sum_{d=1}^{D} w_d z_{do} \\
& \text{s.t.} \sum_{r=1}^{s} u_r y_{ro} = e_o^{\text{centralized}} \\
& \sum_{d=1}^{D} w_d z_{dj} - \sum_{i=1}^{m} v_i x_{ij} \le 0 \quad j = 1,2,\ldots,n \\
& \sum_{r=1}^{s} u_r y_{rj} - \sum_{d=1}^{D} w_d z_{dj} \le 0 \quad j = 1,2,\ldots,n \\
& \sum_{i=1}^{m} v_i x_{io} = 1 \\
& w_d \ge 0, \quad d = 1,2,\ldots,D; \quad v_i \ge 0, \quad i = 1,2,\ldots,m; \quad u_r \ge 0, r = 1,2,\ldots,s. \qquad (12.5)
\end{aligned}
$$

This yields the minimum of $e_o^{2,\text{centralized}}$, namely, $e_o^{2-} = e_o^{\text{centralized}} / e_o^{1+}$. The maximum of $e_o^{2,\text{centralized}}$ can be calculated via the following linear program:

$$
\begin{aligned}
& e_o^{2+} = \text{Max} \sum_{r=1}^{s} u_r y_{ro} \\
& \text{s.t.} \sum_{r=1}^{s} u_r y_{ro} - e_o^{\text{centralized}} \cdot \sum_{i=1}^{m} v_i x_{io} = 0 \\
& \sum_{r=1}^{s} u_r y_{rj} - \sum_{d=1}^{D} w_d z_{dj} \le 0, \quad j = 1,2,\ldots,n \\
& \sum_{d=1}^{D} w_d z_{dj} - \sum_{i=1}^{m} v_i x_{ij} \le 0, \quad j = 1,2,\ldots,n \\
& \sum_{d=1}^{D} w_d z_{do} = 1 \\
& w_d \ge 0, \quad d = 1,2,\ldots,D; \quad v_i \ge 0, \quad i = 1,2,\ldots,m; \quad u_r \ge 0, \quad r = 1,2,\ldots,s, \qquad (12.6)
\end{aligned}
$$

and the minimum of $e_o^{1,\text{centralized}}$ is then calculated as $e_k^{1-} = e_o^{\text{centralized}}/e_o^{2+}$ Note that $e_o^{1-} = e_o^{1+}$ if and only of $e_o^{2-} = e_o^{2+}$. Note also if $e_o^{1-} = e_o^{1+}$ or $e_o^{2-} = e_o^{2+}$, then $e_o^{1,\text{centralized}}$ and $e_o^{2,\text{centralized}}$ are uniquely determined via model (12.3). If $e_o^{1-} \neq e_o^{1+}$ or $e_o^{2-} \neq e_o^{2+}$, Liang et al. (2008) develop a procedure to obtain an alternative decomposition of $e_o^{1,\text{centralized}}$ and $e_o^{2,\text{centralized}}$.

12.4 Stackelberg Game

In the previous section, we examined the cooperative or centralized game approach to the two-stage problem. In this section, we look at the two-stage process from the perspective of the noncooperative game. The noncooperative approach is characterized by the leader–follower Stackelberg game. For example, consider a case of a supply chain where there is noncooperative advertising on the part of the manufacture (leader) and the retailer (follower). The manufacturer determines its optimal brand name investment and local advertising allowance based on an estimation of the local advertisement by the retailer to maximize its profit. The retailer, as a follower on the other hand, based on the information from the manufacturer, finds out the optimal local advertisement cost to maximize its profit (Huang and Li 2001).

In a similar manner, if we assume that the first stage is the leader, then the first stage performance is more important, and the efficiency of the second stage is computed subject to the requirement that the efficiency of the first stage is to stay fixed. We first calculate the efficiency for the first stage. Based upon the CRS model, we have for a specific DMU_o

$$
\begin{aligned}
& e_o^{1*} = \text{Max} \sum_{d=1}^{D} w_d z_{do} \\
& \text{s.t.} \sum_{d=1}^{D} w_d z_{dj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0, \quad j = 1,2,\ldots,n \\
& \sum_{i=1}^{m} v_i x_{io} = 1 \\
& w_d \geq 0, \quad d = 1,2,\ldots,D; \qquad v_i \geq 0, \quad i = 1,2,\ldots,m.
\end{aligned} \tag{12.7}
$$

Note that model (12.7) is in fact the standard CRS DEA model, i.e., e_o^{1*} is the regular DEA efficiency score.

Once we obtain the efficiency for the first stage, the second stage will only consider w_d that maintains $e_o^1 = e_o^{1*}$. Or, in other words, the second stage now treats $\sum_{d=1}^{D} w_d z_{dj}$ as the "single" input subject to the restriction that the efficiency score of the first stage remains at e_o^{1*}. The model for computing e_o^2, the second stage's efficiency, can be calculated as (Liang et al. (2008))

$$
\begin{aligned}
e_o^{2*} &= \text{Max} \frac{\sum_{r=1}^{s} U_r y_{ro}}{Q \sum_{d=1}^{D} w_d z_{do}} \\
\text{s.t.} & \frac{\sum_{r=1}^{s} U_r y_{rj}}{Q \sum_{d=1}^{D} w_d z_{dj}} \le 1, \qquad j = 1,2,\ldots,n \\
& \sum_{d=1}^{D} w_d z_{dj} - \sum_{i=1}^{m} v_i x_{ij} \le 0, \qquad j = 1,2,\ldots,n \\
& \sum_{i=1}^{m} v_i x_{io} = 1 \\
& \sum_{d=1}^{D} w_d z_{do} = e_o^{1*} \\
& U_r, Q, w_d, v_i \ge 0, \quad r = 1,2,\ldots,s; \qquad d = 1,2,\ldots,D; \quad i = 1,2,\ldots,m
\end{aligned} \tag{12.8}
$$

Note that in model (12.8), the efficiency of the first stage is set equal to e_o^{1*}. Let $u_r = U_r/Q, \quad r = 1,2,\ldots,s$. Model (12.8) is then equivalent to the following linear model:

$$
\begin{aligned}
e_o^{2*} &= \text{Max} \sum_{r=1}^{s} u_r y_{ro} / e_o^{1*} \\
\text{s.t.} & \sum_{r=1}^{s} u_r y_{rj} - \sum_{d=1}^{D} w_d z_{dj} \le 0, \qquad j = 1,2,\ldots,n \\
& \sum_{d=1}^{D} w_d z_{dj} - \sum_{i=1}^{m} v_i x_{ij} \le 0, \qquad j = 1,2,\ldots,n \\
& \sum_{i=1}^{m} v_i x_{io} = 1 \\
& \sum_{d=1}^{D} w_d z_{do} = e_o^{1*} \\
& w_d \ge 0, \quad d = 1,2,\ldots,D; \quad v_i \ge 0, \quad i = 1,2,\ldots,m; \quad u_r \ge 0, \quad r = 1,2,\ldots,s.
\end{aligned} \tag{12.9}
$$

In a similar manner, if we assume the second stage as the leader, we then calculate the regular DEA efficiency ($e_o^{2^o}$) for the second stage first using the CCR model. Once we obtain the second stage efficiency, the efficiency for the first stage, namely $e_o^{1^o}$, is calculated via the following linear program (see Liang et al. 2008):

$$\frac{1}{e_o^{1o}} = \text{Min} \sum_{i=1}^{m} v_i x_{io}$$

$$\text{s.t.} \sum_{d=1}^{D} w_d z_{dj} - \sum_{i=1}^{m} v_i x_{ij} \le 0, \qquad j = 1,2,\ldots,n$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{d=1}^{D} w_d z_{dj} \le 0, \qquad j = 1,2,\ldots,n$$

$$\sum_{d=1}^{D} w_d z_{do} = 1$$

$$\sum_{r=1}^{s} u_r y_{ro} = e_o^{2o}$$

$$w_d \ge 0, \;\; d = 1,2,\ldots,D; \quad v_i \ge 0, \;\; i = 1,2,\ldots,m; \quad u_r \ge 0, \;\; r = 1,2,\ldots,s. \tag{12.10}$$

We note that in (12.9), $e_o^{1*} \cdot e_o^{2*} = \sum_{r=1}^{s} u_r^* y_{ro}$ at optimality, with $\sum_{i=1}^{m} v_i^* x_{io} = 1$, i.e., $e_o^{1*} \cdot e_o^{2*} = \sum_{r=1}^{s} u_r^* y_{ro} / \sum_{i=1}^{m} v_i^* x_{io}$. Note also that at optimality, $\sum_{r=1}^{s} u_r^* y_{ro} / \sum_{i=1}^{m} v_i^* x_{io} = e_o^{1o} \cdot e_o^{2o}$ in model (12.10). This indicates that the leader–follower approach also implies an efficiency decomposition for the two-stage process, i.e., the overall efficiency is a product of efficiencies of individual stages. Further, note that in the first-stage leader case, e_o^{1*} and e_o^{2*}, and in the second-stage leader case, e_o^{1o} and e_o^{2o}, are optimal values to linear programs. Therefore, such efficiency decomposition is unique, and is not affected by possible multiple optimal solutions. However, the two approaches may not yield the same efficiency decomposition.

Note that ultimately, a common set of weights is used at both stages in both centralized and Stackelberg game approaches. However, in the Stackelberg game approach, the efficiency scores of two stages, e_o^1 and e_o^2, are not optimized simultaneously.

Liang et al. (2008) also study the relationships among noncooperative and centralized models and the standard DEA approach. We here summarize their findings.

Let θ_o^1 and θ_o^2 be the standard CRS efficiency scores for the two stages.

Theorem 1. If there is only one intermediate measure, then $e_o^{1*} = \theta_o^1$ and $e_o^{2*} = \theta_o^2$ regardless of the assumption of whether the first stage is a leader or follower, where e_o^{1*} and e_o^{2*} are obtained via the noncooperative approach.

Theorem 1 indicates that when there is only one intermediate measure, the noncooperative approach yields the same result as applying the standard DEA model to each stage.

Under the condition of multiple intermediate measures, we have

Theorem 2. For a specific DMU_o, $e_o^{\text{centralized}} \geq e_o^{1*} \cdot e_o^{2*}$, where $e_o^{\text{centralized}}$ is the optimal value to model (12.3), and e_o^{1*} and e_o^{2*} are obtained via the noncooperative (leader–follower) approach.

Based upon Theorems 1 and 2, we must have

Theorem 3. If there is only one intermediate measure, then $e_o^{\text{centralized}} = \theta_o^1 \cdot \theta_o^2$ with $\theta_o^1 = e_o^{1,\text{centralized}}$ and $\theta_o^2 = e_o^{2,\text{centralized}}$, where θ_o^1 and θ_o^2 are the CRS efficiency scores for the two stages, respectively, and $e_o^{1,\text{centralized}}$ and $e_o^{2,\text{centralized}}$ are defined in (12.4).

When there is only one intermediate measure, Theorem 3 indicates that (1) both the noncooperative and centralized models yield the same result as applying the standard DEA model to each stage and (2) the efficiency decomposition is unique.

We finally note that the following is true with respect to the relations between the noncooperative and centralized approaches.

Theorem 4.

1. $e_o^{1,\text{centralized}} \geq e_o^{1*}$ and $\theta_o^2 (= e_o^{2*}) \geq e_o^{2,\text{centralized}}$ when the second stage is the leader,
2. $e_o^{2,\text{centralized}} \geq e_o^{2*}$ and $\theta_o^1 (= e_o^{1*}) > e_o^{1,\text{centralized}}$ when the first stage is the leader.

12.5 DEA Model for General Multistage Serial Processes Via Additive Efficiency Decomposition

We now consider the P-stage process pictured in Fig. 12.2. We denote the input vector to stage 1 by z_o. The output vectors from stage p ($p = 1, \ldots, P$) take two forms, namely z_p^1 and z_p^2. Here, z_p^1 represents that output that leaves the process at this stage and is not passed on as input to the next stage. The vector z_p^2 represents the amount of output that becomes input to the next $(p + 1)$ stage. These types of intermediate measures are called *links* in Tone and Tsutsui (2009).

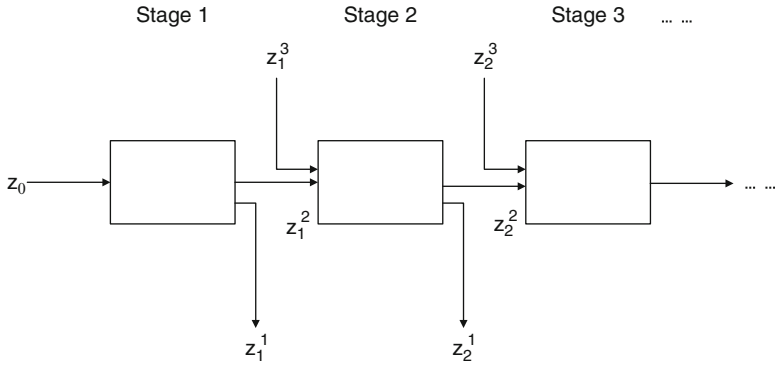


Fig. 12.2 Serial multistage DMU

In addition, there is the provision for new inputs z_p^3 to enter the process at the beginning of stage $p + 1$. Specifically, when $p = 2, 3, \ldots$, we define

1. z_{pr}^{j1} the rth component ($r = 1, \ldots, R_p$) of the R_p-dimensional *output* vector for DMU_j flowing from stage p, that *leaves* the process at that stage, and is not passed on as an input to stage $p + 1$.
2. z_{pk}^{j2} the kth component ($k = 1, \ldots, S_p$) of the S_p-dimensional *output* vector for DMU_j flowing from stage p, and is passed on as a portion of the *inputs* to stage $p + 1$.
3. $z_{pi}^{j3}$the ith component ($i = 1, \ldots, I_p$) of the I_p-dimensional *input* vector for DMU_j at the stage $p + 1$, that enters the process at the beginning of that stage.

Note that in the last stage P, all the outputs are viewed as z_{pr}^{j1}, as they leave the process.

We denote the multipliers (weights) for the above factors as

1. u_{pr} is the multiplier for the output component z_{pr}^{j1} flowing from stage p.
2. η_{pk} is the multiplier for the output component z_{pk}^{j2}at stage p, and is as well the multiplier for that same component as it becomes an input to stage $p + 1$.
3. v_{pi} is the multiplier for the input component z_{pi}^{j3} entering the process at the beginning of stage $p + 1$.

Thus, when $p = 2, 3, \ldots$, the efficiency ratio for DMU_j (for a given set of multipliers) would be expressed as follows:

$$\theta_p = \left(\sum_{r=1}^{R_p} u_{pr} z_{pr}^{j1} + \sum_{k=1}^{S_p} \eta_{pk} z_{pk}^{j2} \right) \Bigg/ \left(\sum_{k=1}^{S_{p-1}} \eta_{p-1k} z_{p-1k}^{j2} + \sum_{i=1}^{I_p} v_{p-1i} z_{p-1i}^{j3} \right) \quad (12.11)$$

Note that there are no outputs flowing into stage 1. The efficiency measure for stage 1 of the process (namely, $p = 1$), for DMU_j becomes

$$\theta_1 = \left(\sum_{r=1}^{R_1} u_{1r} z_{1r}^{j1} + \sum_{k=1}^{S_1} \eta_{1k} z_{1k}^{j2} \right) \Bigg/ \sum_{i=1}^{I_0} v_{0i} z_{0i}^{j} \quad (12.12)$$

where z_{0i}^{j} are the (only) inputs to the first stage represented by the input vector z_o.

We claim that the overall efficiency measure of the multistage process can reasonably be represented as a convex linear combination of the P (stage-level) measures, namely

$$\theta = \sum_{p=1}^{P} w_p \theta_p \quad \text{where} \quad \sum_{p=1}^{P} w_p = 1.$$

Note that the weights w_p are intended to represent the relative importance or contribution of the performances of individual stages p to the overall performance of the entire process. One reasonable choice for weights w_p is the proportion of total resources for the process that are devoted to stage p, and reflecting the relative size

of that stage. To be more specific, $\sum_{i=1}^{I_0} v_{0i} z_{0i}^{j} + \sum_{p=2}^{P} \left(\sum_{k=1}^{S_{p-1}} \eta_{p-1k} z_{p-1k}^{j2} + \sum_{i=1}^{I_p} v_{p-1i} z_{p-1i}^{j3} \right)$ represents the total size of or total amount of resources consumed by the entire process, and we define the w_p to be the proportion of the total input used at the pth stage. We then have

$$w_1 = \sum_{i=1}^{I_0} v_{0i} z_{0i}^{j} \Bigg/ \left\{ \sum_{i=1}^{I_0} v_{0i} z_{0i}^{j} + \sum_{p=2}^{P} \left(\sum_{k=1}^{S_{p-1}} \eta_{p-1k} z_{p-1k}^{j2} + \sum_{i=1}^{I_p} v_{p-1i} z_{p-1i}^{j3} \right) \right\},$$

$$w_p = \left(\sum_{k=1}^{S_{p-1}} \eta_{p-1k} z_{p-1k}^{j2} + \sum_{i=1}^{I_p} v_{p-1i} z_{p-1i}^{j3} \right) \Bigg/ \left\{ \sum_{i=1}^{I_0} v_{0i} z_{0i}^{j} + \sum_{p=2}^{P} \left(\sum_{k=1}^{S_{p-1}} \eta_{p-1k} z_{p-1k}^{j2} + \sum_{i=1}^{I_p} v_{p-1i} z_{p-1i}^{j3} \right) \right\}., \quad p>1.$$

Thus, we can write the overall efficiency θ in the form

$$\theta = \sum_{p=1}^{P} \left(\sum_{r=1}^{R_p} u_{pr} z_{pr}^{j1} + \sum_{k=1}^{S_p} \eta_{pk} z_{pk}^{j2} \right) \Bigg/ \left\{ \sum_{i=1}^{I_0} v_{0i} z_{0i}^{j}. + \sum_{p=2}^{P} \left(\sum_{k=1}^{S_{p-1}} \eta_{p-1k} z_{p-1k}^{j2} + \sum_{i=1}^{I_p} v_{p-1i} z_{p-1i}^{j3} \right) \right\}.$$

We then set out to optimize the overall efficiency θ of the multistage process, subject to the restrictions that the individual measures θ_p must not exceed unity, or in the linear programming format, after making the usual Charnes and Cooper transformation,

$$\begin{aligned}
&\max \sum_{p=1}^{P} \left(\sum_{r=1}^{R_p} u_{pr} z_{pr}^{o1} + \sum_{k=1}^{S_p} \eta_{pk} z_{pk}^{o2} \right) \\
&\text{s.t.} \left\{ \sum_{i=1}^{I_0} v_{0i} z_{0i}^{o}. + \sum_{p=2}^{P} \left(\sum_{k=1}^{S_{p-1}} \eta_{p-1k} z_{p-1k}^{o2} + \sum_{i=1}^{I_p} v_{p-1i} z_{p-1i}^{o3} \right) \right\} = 1 \\
&\left(\sum_{r=1}^{R_1} u_{1r} z_{1r}^{j1} + \sum_{k=1}^{S_1} \eta_{1k} z_{1k}^{j2} \right) \leq \sum_{i=1}^{I_0} v_{0i} z_{0i}^{j} \\
&\left(\sum_{r=1}^{R_p} u_{pr} z_{pr}^{j1} + \sum_{k=1}^{S_p} \eta_{pk} z_{pk}^{j2} \right) \leq \left(\sum_{k=1}^{S_{p-1}} \eta_{p-1k} z_{p-1k}^{j2} + \sum_{i=1}^{I_p} v_{p-1i} z_{p-1i}^{j3} \right) \forall j \\
&u_{pr}, \eta_{pk}, v_{pi}, v_{0i} \geq 0.
\end{aligned} \tag{12.14}$$

Note that we should impose the restriction that the overall efficiency scores for each j should not exceed unity, but since these are redundant, this is unnecessary.

Note again that the w_p, as defined above, are variables related to the inputs and the intermediate measures. By virtue of the optimization process, it can turn out that some $w_p = 0$ at optimality. To overcome this problem, one can impose bounding restrictions $w_p > \beta$, where β is a selected constant.

12.6 General Multistage Processes

In the process discussed in the previous section, it is assumed that the components of a DMU are arranged in series as depicted in Fig. 12.2. There, at each stage p, the inputs took one of two forms, namely (1) those that are outputs from the previous stage p-1 and (2) new inputs that enter the process at the start of stage p. On the output side, those (outputs) emanating from stage p take two forms as well, namely (1) those that leave the system as finished "products" and (2) those that are passed on as inputs to the *immediate* next stage $p + 1$.

The model presented to handle such strict serial processes is easily adapted to more general network structures. Specifically, the efficiency ratio for an overall process can be expressed as the weighted average of the efficiencies of the individual components. The efficiency of any given component is the ratio of the total output to the total input corresponding to that component. Again, the weight w_p to be applied to any component p is expressed as

$$w_p = (\text{component } p \text{ input})/(\text{total input across all components}).$$

There is no convenient way to represent a network structure that would lend itself to a generic mathematical representation analogous to model (12.14) above. The sequencing of activities and the source of inputs and outputs for any given component will differ from one type of process to another. However, as a simple illustration, consider the following two examples of network structures.

12.6.1 Parallel Processes

Consider the process depicted in Fig. 12.3. Here, an initial input vector z_0 enters component 1. Three output vectors exit this component, that is z_1^1 leaves the process, z_1^2 is passed on as an input to component 2, and z_1^3 as an input to component 3. Additional inputs z_1^4 and z_1^5 enter components 2 and 3 respectively, from outside the process. Components 2 and 3 have z_2^1 and z_3^1, respectively as output vectors which are passed on as inputs to component 4, where a final output vector z_4^1 is the result.

Component Efficiencies

Component 1 efficiency ratio: $\theta_1 = \left(u_1 z_1^1 + \eta_1^2 z_1^2 + \eta_1^3 z_1^3\right)/v_0 z_0$

Component 2 efficiency ratio: $\theta_2 = \eta_2^1 z_2^1/\left(\eta_1^2 z_1^2 + v_1 z_1^4\right)$

Component 3 efficiency ratio: $\theta_3 = \eta_3^1 z_3^1/\left(\eta_1^3 z_1^3 + v_2 z_1^5\right)$

Component 4 efficiency ratio: $\theta_4 = u_4 z_4^1/\left(\eta_2^1 z_2^1 + \eta_3^1 z_3^1\right)$

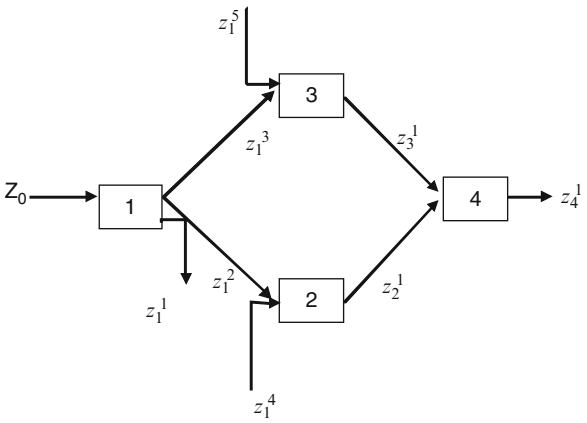


Fig. 12.3 Multistage DMU with parallel processes

Component Weights

Note that the total (weighted) input across all components is given by the sum of the denominators of θ_1 through θ_4, namely

$$I = v_0 z_0 + \eta_1^2 z_1^2 + v_1 z_1^4 + \eta_1^3 z_1^3 + v_2 z_1^5 + \eta_2^1 z_2^1 + \eta_3^1 z_3^1.$$

Now express the w_p as

$$\begin{aligned} w_1 &= v_0 z_0 / I \\ w_2 &= (\eta_1^2 z_1^2 + v_1 z_1^4)/I \\ w_3 &= (\eta_1^3 z_1^3 + v_2 z_1^5)/I \\ w_4 &= (\eta_2^1 z_2^1 + \eta_3^1 z_3^1)/I \end{aligned}$$

With this, the overall network efficiency ratio is given by

$$\theta = \sum_{p=1}^{4} w_p \theta_p = (u_1 z_1^1 + \eta_1^2 z_1^2 + \eta_1^3 z_1^3 + \eta_2^1 z_2^1 + \eta_3^1 z_3^1 + u_4 z_4^1)/I,$$

and one then proceeds, as in (12.14) above, to derive the efficiency of each DMU and its components.

12.6.2 Nonimmediate Successor Flows

In the previous example all flows of outputs from a stage or component either leave the process entirely or enter as an input to an *immediate successor* stage. In Fig. 12.2, stage p outputs flow to stage $p + 1$. In Fig. 12.3, the same is true except that there is more than one immediate successor of stage 1.

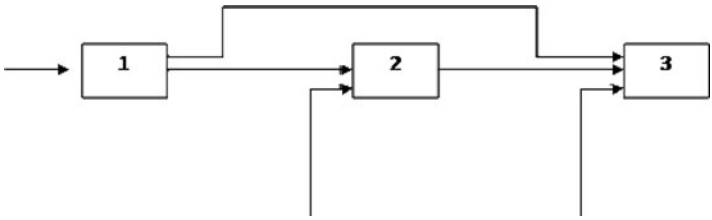


Fig. 12.4 Nonimmediate successor flows

Consider Fig. 12.4. Here, the inputs to stage 3 are of three types, namely outputs from stage 2, inputs coming from outside the process, and outputs from a previous, but not immediately previous stage. Again the above rationale for deriving weights w_p. can be applied and a model equivalent to (12.14) solved to determine the decomposition of an overall efficiency score into scores for each of the components in the process.

12.7 Conclusions

The current chapter reviews some of the literature on DEA models for measuring the performance of DMUs with two-stage process structures. The focus is on those particular models that can lead to an efficiency decomposition of the overall efficiency of the entire process. We have shown that many of the existing approaches, network DEA, Kao and Hwang (2008), Liang et al. (2006), and Liang et al. (2008) are all centralized or cooperative models. The Stackelberg or leader–follower model of Liang et al. (2008) provides a useful alternative approach.

We note that for systems composed of a set of processes connected in parallel, Kao (2009a) develops a DEA model to decompose the inefficiency slack of the system to the sum of those of the component processes. Further, Kao (2009b) provides a mathematical relationship between the system efficiency and the process efficiencies under network DEA. One alternative to these approaches is to apply the Stackelberg game concept as shown in Liang et al. (2008).

While the approach of Liang et al. (2008) focuses on the geometric (multiplicative) efficiency decomposition of the overall efficiency, Chen et al. (2009) and Cook et al. (2010b) develop an approach for additive efficiency decomposition. Unlike the nonlinear model in Liang et al. (2006), their model is not only linear, but also can be applied to variable returns to scale (VRS) situations. Note that because of the extra free-in-sign variable in the VRS DEA model, the use of geometric mean of the efficiency scores of the two individual stages will render the resulting DEA model highly nonlinear. Therefore, it appears that efficiency decomposition in an additive form is a reasonable and computationally tractable way to measure the performance of two-stage processes or network structures under VRS. So far, the existing approaches for other DMU structures, e.g., network DEA approach

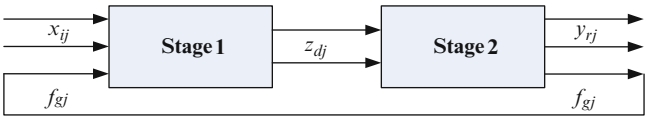


Fig. 12.5 Two-stage process with feedback

and the approach of Castelli et al. (2004) are all developed under the assumption of CRS, which, it would appear, cannot be directly modified to fit the VRS assumptions.

In addition to the above-mentioned problems arising from two-stage processes, there is the previously mentioned concern that none of the DMUs may turn out to be efficient. Further research is needed to investigate this issue.

Cook et al. (2010b) develop a DEA approach for DMUs that have a general multistage or network structure. They first examine pure serial networks where each stage has its own inputs and two types of outputs. One type of output from any given stage p is passed on as an input to the next stage, and the other type exits the process at stage p. In general, the intermediate measures are those that exist between two members of the network. In many cases, the intermediate measures are obvious, as indicated in our examples mentioned in Sect. 12.1. Tone and Tsutsui (2009) provide other good examples. Sometimes, the selection of intermediate measures is not so obvious. The important thing is that intermediate measures are neither "inputs" (to be reduced) nor "outputs" (to be increased), rather these measures need to be "coordinated" to determine their efficient levels (see Kao and Hwang 2008; Liang et al. 2008.)

Finally, Liang et al. (2011) extend the above approaches to include those situations where outputs from the second stage can be fed back as inputs to the first stage (see Fig. 12.5). Such feed-back variables thus serve a dual role.

References

Castelli L, Pesenti R, Ukovich W. DEA-like models for the efficiency evaluation of hierarchically structured units. Eur J Oper Res. 2004;154(2):465–76.

Castelli L, Pesenti R, Ukovich W. A classification of DEA models when the internal structure of the decision making units is considered. Ann Oper Res. 2008;173:207–35.

Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res. 1978;2:429–44.

Chen C-M. A network-DEA model with new efficiency measures to incorporate the dynamic effect in production networks. Eur J Oper Res. 2009;194(3):687–99.

Chen Y, Zhu J. Measuring information technology's indirect impact on firm performance. Inf Tech Manage J. 2004;5(1–2):9–22.

Chen Y, Liang L, Yang F, Zhu J. Evaluation of information technology investment: a data envelopment analysis approach. Comput Oper Res. 2006;33(5):1368–79.

Chen Y, Cook WD, Li N, Zhu J. Additive efficiency decomposition in two-stage DEA. Eur J Oper Res. 2009;196:1170–6.

Chilingerian J, Sherman HD. Health care applications: from hospitals to physician, from productive efficiency to quality frontiers. In: Cooper WW, Seiford LM, Zhu J, editors. Handbook on data envelopment analysis. Boston: Springer; 2004.

Cook WD, Liang L, Zhu J. Measuring performance of two-stage network structures by DEA: a review and future perspective. Omega. 2010a;38:423–30.

Cook WD, Zhu J, Bi G, Yang F. Network DEA: additive efficiency decomposition. Eur J Oper Res. 2010b;207(2):1122–9.

Färe R, Whittaker G. An intermediate input model of dairy production using complex survey data. J Agric Econ. 1995;46(2):201–13.

Färe R, Grosskopf S. Productivity and intermediate products: a frontier approach. Econ Lett. 1996;50:65–70.

Gaski JF. The theory of power and conflict in channels of distribution. J Mark. 1984;15:107–11.

Huang ZM, Li SX. Co-Op advertising models in a manufacturing-retailing supply chain: a game theory approach. Eur J Oper Res. 2001;135:527–44.

Kao C. Efficiency measurement for parallel production systems. Eur J Oper Res. 2009a;196 (3):1107–12.

Kao C. Efficiency decomposition in network DEA: a relational model. Eur J Oper Res. 2009b;192:949–62.

Kao C, Hwang SN. Efficiency decomposition in two-stage data envelopment analysis: an application to non-life insurance companies in Taiwan. Eur J Oper Res. 2008;185(1):418–29.

Liang L, Cook WD, Zhu J. DEA models for two-stage processes: game approach and efficiency decomposition. Nav Res Logist. 2008;55:643–53.

Liang L, Yang F, Cook WD, Zhu J. DEA models for supply chain efficiency evaluation. Ann Oper Res. 2006;145(1):35–49.

Liang L, Li ZQ, Cook WD, Zhu J. DEA efficiency in two-stage networks with feed back. IIE Trans. 2011;43:309–22.

Seiford LM, Zhu J. Profitability and marketability of the top 55 US commercial banks. Manag Sci. 1999;45(9):1270–88.

Sexton TR, Lewis HF. Two-stage DEA: an application to major league baseball. J Product Anal. 2003;19(2–3):227–49.

Tone K, Tsutsui M. Network DEA: a slacks-based measure approach. Eur J Oper Res. 2009;197(1):243–52.

Wang CH, Gopal R, Zionts S. Use of data envelopment analysis in assessing information technology impact on firm performance. Ann Oper Res. 1997;73:191–213.

Zhu J. Multi-factor performance measure model with an application to Fortune 500 companies. Eur J Oper Res. 2000;123(1):105–24.

Chapter 13
Assessing Bank and Bank Branch Performance

Modeling Considerations and Approaches

Joseph C. Paradi, Zijiang Yang, and Haiyan Zhu

Abstract The banking industry has been the object of DEA analyses by a significant number of researchers and probably is the most heavily studied of all business sectors. Various DEA models have been applied in performance assessing problems, and the banks' complex production processes have further motivated the development and improvement of DEA techniques. The main application areas for DEA in bank and branch performance analysis include the following: efficiency ranking; resource allocation, efficiency trends investigation; environmental impacts compensation; examining the impacts of new technology, ownership, deregulation, corporate, economic, and political events, etc.

Keywords Banking performance • DEA models • Performance • Efficiency

13.1 Introduction

As the principal sources of financial intermediation and means of making payments, banks play a vital role in a country's economic development and growth. In addition to their large economic significance, the existence of an increasingly competitive market structure highlights the importance of evaluating the banks' performance to continuously improve their functions and monitor their financial condition. There are many uses for performance analyses by bank regulators who need to determine how the industry will respond to the introduction of new regulations, nontraditional entrants and distribution systems, worldwide competition, and thereafter, to direct future government policies. Equally important use of such analyses is made by bank management concerned about the effectiveness of their resource allocation, the

J.C. Paradi (✉)
Centre for Management of Technology and Entrepreneurship,
University of Toronto, Toronto, ON, Canada M5S 3E5
e-mail: paradi@mie.utoronto.ca

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_13,

impact of ongoing structural changes on bank operations, and their ability to realign their businesses with the current and most profitable trends. The challenge still remains in selecting the most suitable methodology for these types of assessments.

Historically, banks have been in the forefront of trying to improve their operating efficiency, but typically they have not had at their disposal a sound analysis system to evaluate their performance at both institutional and branch network levels. Capturing the essential aspects of the operating process, such a system should yield a relevant and trustworthy measure, suitable to establish benchmarks. Indicative of the unit's ability to use its resources to generate desirable outcomes, this measure should lead to a better understanding of the process in terms of what is achieved and how it is achieved. It should allow a meaningful investigation of hypotheses concerning the sources of inefficiency with considerations of internal and external environmental impacts. Finally, it should provide management with the tools with which to monitor performance.

The events of the worldwide financial meltdown in 2008–2009 are sure to result in more regulatory demands for reporting by banks. Moreover, operational issues and policies are under increased scrutiny in an attempt to prevent another series of such problems. There are many reasons for the growth of money laundering, and this is another area of concern for banks. New approaches will be required to respond to these new challenges.

13.2 Performance Measurement Approaches in Banking

Long before DEA was introduced, banks have been engaged in performance measurement techniques, and most large banks have from one person to whole departments engaged in continuous performance monitoring activities. However, the arsenal of tools and the variety of approaches are typically limited to a few relatively simple methods.

13.2.1 Ratio Analysis

Ratio analysis has historically been the standard technique used by regulators, industry analysts, and management to examine banking performance. Ratios measure the relationship between two variables chosen to provide insights into different aspects of the banks' multifaceted operations, such as liquidity, profitability, capital adequacy, asset quality, risk management, and many others. Any number of ratios can be designed depending on the objective of the analysis, generally for comparisons within the same bank over different time periods, as well as for benchmarking with reference to other banks. Key Performance Indicators (KPIs) are commonly used by investors and financial institutions to study the banks'

business performance and financial condition. Various KPIs have been designed to provide a high-level snapshot of the organization from different perspectives.

Compustat/Research Insight, for instance, offers a number of corporate and financial information research databases of publicly traded corporations compiled by Standard and Poors (http://www.compustat.com). This financial analysis software package contains up to 20 years of financial ratios, stock market prices, geographic segment data, and S&P's ratings and rankings.

Although the traditional ratio measures are attractive to analysts due to their simplicity and ease of understanding, there are several limitations that must be considered. For example, the analysis assumes comparable units, which implies constant returns-to-scale (Smith 1990). Each of the indicators yields a one-dimensional measure by examining only a part of the organization's activities, or combining the multiple dimensions into a single, unsatisfactory number. Moreover, the seemingly unlimited numbers of ratios that can be created from financial statement data are often contradictory and confusing, and thus ineffective for the assessment of overall performance. This overly simplistic analytical approach offers no objective means of identifying inefficient units and requires a biased separation of the inefficient and efficient levels. Failure to account for multidimensional input and output processes, combined with its inability to pinpoint the best performers in any culturally homogeneous group, makes ratio analysis inadequate for efficiency evaluations.

13.2.2 Frontier Efficiency Methodologies

The limitations associated with ratio analysis have led to the development of more advanced tools for assessing corporate performance. In recent years, research on the performance of financial institutions has increasingly focused on the production frontier based models, which estimate how well a firm performs relative to the best firms if they are doing business under the same operating conditions. These best firms are identified from the dataset, and they form the empirically efficient frontier. The main advantage of frontier efficiency over other indicators of performance is that it offers overall objective numerical efficiency scores with economic optimization mechanisms in complex operational environments (Berger and Humphrey 1997). In addition, management is provided with a framework that supports the planning, decision-making and control processes. Sources and magnitude of inefficiency can be determined that may ultimately lead to a reduction in the cost of operations or an increase in the services provided without expending additional resources. Achievable targets for inefficient units and the effects of environmental variables can be determined to provide additional insights and improve the overall understanding of production systems.

In the past three decades, five popular frontier efficiency approaches have been used in bank efficiency measurements; three of them are parametric econometric approaches: stochastic frontier approach (SFA), thick frontier

approach (TFA), and distribution-free approach (DFA). The other two are nonparametric linear programming approaches: data envelopment analysis (DEA) and free disposal hull (FDH). These approaches primarily differ in the assumptions imposed on the specifications of the efficient frontier, the existence of random error, and the distribution of the inefficiencies and random error (Bauer 1990; Berger and Humphrey 1997). Econometric analyses require an a priori specification of the form of the production function and typically include two error components: an error term that captures inefficiency and a random error. While mathematical, nonparametric methods require few assumptions when specifying the best-practice frontier, they generally do not account for random errors. Some researchers (such as Ferrier and Lovell 1990; Resti 1997; Bauer et al. 1998; Weill 2004) have comparatively studied the consistency and robustness of the estimations generated by the various frontier techniques.

13.2.3 Other Performance Evaluation Methods

Other performance evaluation methods include the following: Multivariate Statistical Analysis (e.g., Zopounidis et al. 1995; Canbas et al. 2005); Analytic Hierarchy Process (e.g., Frei and Harker 1999; Seçme et al. 2009); Grey Relation Analysis (e.g., Ho and Wu 2006); Balanced Scorecard (e.g., Kim and Davidson 2004; Wu et al. 2009). Each method has its own advantages and disadvantages.

13.3 Data Envelopment Analysis in Banking

Sherman and Gold (1985) wrote the first significant DEA bank analysis paper and started what turned out to be a long list of DEA applications to banking from several different angles:

- Countrywide bank (companies) analysis
- Cross-national banking analysis
- Bank merger efficiencies
- Bank branch analysis within one banking organization
- Branch deployment strategies
- Internal performance evaluation along the production process
- Customer service quality analysis
- Project selection and planning
- Staff allocation efficiency
- Personnel retention and work intensity targets

Unparalleled growth in the number of theoretical and empirical investigations of banking efficiency has resulted in the main operational research journals devoting special issues just to this research area (i.e., Journal of Productivity Analysis

(1993), Journal of Banking and Finance (1993), European Journal of Operational Research (1997), Annals of Operations Research (1997, 2006), Interfaces (1999), International Journal of Information Technology and Decision Making (2005), and The Journal of the Operational Research Society (2009)). A considerable number of papers have been published on the banking industry (both banks as entities and branches) using DEA since the technology was introduced. Berger and Humphrey (1997) summarized these works and listed 41 DEA studies in a number of countries; US banking research predominated with 23 papers. Fethi and Pasiouras (2010) review the 196 studies that employ operational research and artificial intelligence techniques in the assessment of bank performance; among them there are 151 studies using DEA-like techniques to estimate various measures of bank efficiency and productivity. From 1997 to 2010, our own survey identifies 225 DEA applications in the banking industry, 162 at the institution level, and 63 at the branch level. These applications cover 43 countries/regions; among them there are 28 studies with international scope. Clearly, the work keeps going on and will likely continue in the foreseeable future, especially the aftermath of the 2008–2009 financial meltdown.

13.3.1 Banking Corporations

Based on our survey, which updated Berger and Humphrey (1997), there have been a considerably larger number of studies that focus on the efficiency of banking institutions at the industry level compared to those that measure performance at the branch level. Lack of publicly available branch data is perhaps the reason behind this situation.

13.3.1.1 In-Country

In any country that has a sufficient number of banks, an in-country study using DEA is quite feasible. A number of such studies have been done, many of them involving the US banking system, while other researchers focused on India, Italy, Japan, Spain, and Sweden. Favero and Papi (1995) derived measures of technical and scale efficiencies in Italy examining a cross section of 174 Italian banks in 1991. Bhattacharyya et al. (1997) examined the productive efficiency of 70 Indian banks during 1986–1991, which was a period of liberalization. They found the publicly owned Indian banks to have been the most efficient, followed by foreign owned banks and privately owned Indian banks. Thompson et al. (1997) applied classification, sensitivity, uniqueness, linked cones and profit ratios to a bank panel of the 100 largest US banks from 1986 to 1991. Tortosa-Ausina (2002) explored the 121 Spanish bank efficiency changes over the period 1985–1995. Drake and Hall (2003) utilized DEA to analyze the technical and scale efficiency in Japanese banking using a cross-sectional sample of 149 banks.

Das and Ghosh (2006) investigated the performance of the Indian commercial banking sector including 98 banks during the post reform period of 1992–2002. Bergendahl and Lindblom (2008) examined 88 Swedish savings banks' customer service efficiency. Al-Sharkas et al. (2008) applied DEA to investigate the impact of mergers and acquisitions on the efficiency of the US banking industry over the period of 1986–2002, the sample including 440 bank mergers. Fukuyama and Weber (2008) estimated the inefficiency and the shadow price of problem loans within 126 Japanese banks.

The main research topics of in-country bank efficiency studies include investigating resources of performance inefficiency, the changes in bank efficiencies over time, the determinants of bank efficiency, the impact of economic reforms, ownership, new technologies, and socioeconomic conditions on the bank efficiency, etc.

13.3.1.2 Cross-Country Studies

Cross-country studies are quite difficult because allowances must be made for a host of natural differences between the nations involved. These include real cultural differences (as opposed to DEA cultural issues), regulatory environments, presence in a common market structure (EU, NAFTA), and the "sophistication" of the banking institutions. For example, Berg et al. (1992) examined the banking community in Norway, Sweden, and Finland. They created individual and then combined frontiers to examine the similarities and differences among the banks. The study by Pastor et al. (1997) covered 429 banks in eight developed countries, and pooled cross-country data to define a common frontier. Lozano-Vivas et al. (2002) investigated the operating efficiency differences of 612 commercial banks across 10 European countries. The environmental factors, which were treated as non-discretionary variables, were incorporated into the DEA model including income per capita, salary per capita, population density, density of demand, income per branch, deposit per branch, branches per capita, and branch density. Seven Caribbean country banks were studied by McEachern and Paradi (2007). This was an unusual approach, as there was only one bank from each country, but they were owned by the same parent Canadian bank. Owing to the very similar management culture in the seven banks, the corporate disparity was removed and that allowed the researchers to examine the effects of the national culture on bank branch productivity. Pasiouras (2008) applied a two-stage DEA model on a sample of 715 banks from 95 countries to examine the impact of regulations and supervision approaches on the banks' efficiency. Thoraneenitiyan and Avkiran (2009) investigated the relationship between postcrisis bank restructuring, country-specific conditions, and bank efficiency in East Asian countries from 1997 to 2001, including 110 banks, using an approach that integrates DEA and stochastic frontier analysis.

13.3.2 Bank Branches

Much of the published literature on bank branch performance utilizes methodologies based on nonparametric frontier estimations, primarily DEA and FDH. Appendix B provides 64 DEA-based bank branch studies from 1997 to 2010, containing information on the country, the number of branches analyzed, and the specifications of the model variables.

Most of these studies examine branch performance by assuming various behavioral objectives.

- In the *production* approach, branches are regarded as using labor and capital to produce deposits and loans.
- The *intermediation* approach captures the process in which deposits are being converted into loans.
- *Profitability* is the measure of how well branches generate profits from their use of labor, assets, and capital.

Some studies investigate impacts of the new technologies on the branches' roles. For example, Cook et al. (2000) measured branch performance that focused on their sales and service functions. Cook et al. (2004) studied the new structures of bank branches and the effect of e-business activities on banking performance.

13.3.2.1 Small Number of Branches

The early studies on branch performance are fairly simple and evaluate efficiency using small sample sizes. Sherman and Gold (1985) used a basic CCR model to analyze the production (operational) efficiency of 14 branches belonging to a US savings bank. This then set the direction for subsequent studies by demonstrating the practical and diverse opportunities associated with the methodology.

Based on the work by Vassiloglou and Giokas (1990), Giokas (1991) estimated the relative efficiencies of 17 branches of a bank in Greece using both DEA and log-linear methodologies. Oral and Yolalan (1990) introduced a DEA measurement model that forced each of the 20 branches in a Turkish bank to compare itself with the global leader. This was the first study to include time-based outputs (activities were given standard times) in the production model, instead of the *number of transactions* that were used in traditional DEA applications. Oral et al. (1992) developed two banking models on the same dataset for analyzing both the operational efficiency and profitability of those Turkish bank branches.

Parkan (1987) used the CCR model to benchmark 35 branches of a Canadian bank. Al-Faraj et al. (1993) applied the basic formulations of DEA to assess the performance of 15 bank branches in Saudi Arabia. They used eight inputs and seven outputs, and subsequently identified all but three branches as relatively efficient. They inadvertently illustrated one of the limitations associated with the DEA approach: its inability to effectively discriminate between efficient and inefficient

units when a limited number of observations, relative to the number of input/output variables, were used. Sherman and Ladino (1995) reported that the implementation of their DEA results in the restructuring process of the 33 branches belonging to a US bank led to actual annual savings of over $6 million. But this study was potentially biased due to an insufficient number of branches. Cook et al. (2000), using only 20 bank branches of a large Canadian bank, offered a methodology for splitting shared resources among branch activities to maximize the aggregate efficiency scores. The performance of 90 commercial branches of a large Canadian bank was examined by Paradi and Schaffnit (2004). The rate of change of provincial GDP (Gross Domestic Product) was used to account for economic environmental differences. Yavas and Fisher (2005) evaluated and ranked the operational performance of 32 US commercial bank branches.

13.3.2.2 Large Number of Branches

Tulkens (1993) examined 773 branches of a public bank and 911 branches of a private bank in Belgium, using the FDH model for the first time in banking. Drake and Howcroft (1994) reported the efficiencies of 190 branches in UK under both the CCR and BCC assumptions. The analysis was taken further to reveal an optimum branch size in terms of the number of staff employed. Schaffnit et al. (1997) specifically focused on personnel performance using a 291 branch segment of a large Canadian bank. Several DEA models with output and input multiplier constraints were developed to measure the operating efficiency in providing both transactions and maintenance services at the branch.

Lovell and Pastor (1997) evaluated the operating efficiency of 545 branch offices based on the target setting procedure employed in a large bank in Spain. Athanassopoulos (1998) conducted a performance assessment of 580 branches of a commercial bank in UK using cost efficiency, then a new dimension in bank branch performance.

Kantor and Maital (1999) integrated an Activity-Based Cost accounting model and a CCR DEA model to measure inefficiency separately for customer services and transactions in 250 Mideast bank branches. Camanho and Dyson (1999) estimated operational efficiencies of 168 branches of a large bank in Portugal. The efficiency–profitability matrix was then employed for a more comprehensive characterization of their performance profile.

Golany and Storbeck (1999) conducted a multiperiod efficiency analysis of 182 selected branches in USA using a DEA model that allowed for the separation of inputs into discretionary and nondiscretionary (i.e., not controllable by management) factors. A resource-allocation model was introduced to estimate the maximal output enhancement that could be expected from the branch given its level of inputs. Zenios et al. (1999) extended the DEA framework to capture the effects of the environment on the efficiency measures of 144 branches in Cyprus. They examined whether superior efficiency was the result of the technology used or managerial practices.

Sherman and Rupert (2006) analyzed operating efficiencies of a 217-branch network formed in a merger of four banks. Each branch was benchmarked against best-practice branches in the combined merged bank as well as best-practice branches within each premerger bank. Paradi et al. (2010a) examined the productivity and profitability of 1,213 branches of two Canadian banks while considering the corporate cultural impacts on branch performance. Deville (2009) analyzed the operational performances of 1,611 French branches that formed 16 regional groups and operated in six different business environments.

13.3.2.3 Branch Studies Incorporating Service Quality

There are mainly two ways to incorporate service quality factors into branch performance analyses, either directly into the DEA model or conducting post hoc analyses on the relationship between the DEA efficiency found and the service quality reported. Golany and Storbeck (1999) incorporated customer loyalty and customer satisfaction into their DEA model as outputs to evaluate the performances of bank branches by seeing the bank as a provider of financial services. Soteriou and Stavrinides (1997, 2000) incorporated service quality as an output to provide suggestions toward internal customer service quality improvements. Soteriou and Zenios (1999) built on this work to provide a more complete assessment of the overall performance of a larger network consisting of 144 branches in Cyprus. The paper attempted to link service quality, operations, and profitability in a common framework of efficiency benchmarks. Sherman and Zhu (2006) developed a multistage DEA model to incorporate quality measures into the DEA efficiency analysis. Athanassopoulos (1997, 2000) conducted post hoc analysis to capture the impacts of service quality on branch operating efficiency. Portela and Thanassoulis (2007) also investigated the links between service quality and branch operating and profit efficiency.

13.3.2.4 Unusual Banking Applications of DEA

Nash and Sterna-Karwat (1996) measured the cross-selling effectiveness of financial products for 75 bank branches in Australia. Soteriou and Zenios (1999) presented a novel approach to examine the efficiency of bank product costing at the branch level. The focus was on allocating total branch costs to the product mix offered by the branch and obtaining a reliable set of cost estimates for these products. Jablonsky et al. (2004) proposed a DEA model for forecasting branch future efficiency bounds based on an interval input–output data from bank management's pessimistic and optimistic predictions. Stanton (2002) investigated the relationship managers' efficiencies in one of Canada's largest banks.

It is difficult to compare different banks' branch networks because the corporate culture of the banks may influence branch performance in different ways, such that cross-firm comparisons are perceived as unfair and inequitable.

About 70 branches of three banks were compared using a pair of handicapping factors by Yang and Paradi (2006) to show that a level playing field could be created to deal with culturally different branches.

In summary, it can be seen that there are many reasons for stimulating the analysts' ingenuity in formulating the appropriate DEA models. Most of the work referred to here address real-life issues that are usually less than ideal (from an analytical or theoretical point of view); hence, the analyst must innovate to use what data is available to come up with the answers required by management.

13.4 Model Building Considerations

There are two crucial considerations the analyst must keep in mind when preparing a study in a banking environment – selecting inputs and outputs and the choice of the DEA technology. These issues are pointed out and the reasons for making the choices are discussed as the chapter progresses.

13.4.1 Approaching the Problem

Analysts and managers must come to common ground about what each model is to accomplish. The researcher must satisfy the needs for specific information that is likely to be useful to the bank (branch) manager and to show the way to better performance. Before bank (branch) efficiency can be measured, a definition of what its business processes are is required. Countless studies have been done to determine accurate ways of measuring bank (branch) efficiency, as detailed above. Kinsella (1980) discussed some of the reasons banks were difficult to measure: they offered multiple products, had complex services (many of which were interdependent), provided some services that were not directly paid for, and had complex government regulations that might affect the way in which services were offered or priced. Given these issues, it becomes obvious that there is no one way of accurately measuring bank (branch) efficiency and that clearly a combined set of metrics is required.

Moreover, for the study to be successful from an operational and implementation point of view, it is necessary that management be involved right from the start. Decision makers must understand all model designs and the selection of measures to use in the models and must see the outcomes as fair and equitable.

13.4.2 Input or Output?

When defining models, often the question arises about what to do with an input or output that should not be minimized or maximized respectively. One example of these is bad loans, which is obviously an output, but it is not desirable to reward the

DMU (Decision Making Unit, a bank or a branch here) for having more bad loans than its peers have. So, how can the problem be solved? Three different approaches have been used in the literature: the first is to leave bad loans as an output but use the inverse value. The second method is to move it to the input side where the lower this value, the better. This seems like a fairly simple decision to make. But for the sake of managerial understanding the inverse value on the output side may be preferred. The third one is to treat bad loans as undesirable output with an assumption of weak disposability, which requires that undesirable outputs can be reduced, but at a cost of fewer desirable outputs produced (Färe and Grosskopf 2003).

Now, how about deposits? On the one hand, it could be argued that the higher the value the better because that shows efficiency in attracting depositors. On the other hand, one could make a case that the lower the deposit value, the better, because the bank is doing more lending with less deposits. This, of course, implies that the bank has sources of funds that are cheaper than deposits. Fortunately, the analyst can have it either way, but also both ways if there are several models being built, depending on what the model is intended to achieve.

13.4.3 Too Few DMUs/Too Many Variables

DEA, as many other methods, requires that there be enough observations to allow good separation and discrimination between DMUs. Several methods can be used to address this problem. One is to increase the number of DMUs by using the Windows Analysis approach, which allows different years' observations to be compared to each other.

Another possibility is to decrease the number of inputs and outputs in the models by creating more than one model where each has a fewer total (inputs + outputs) number of variables. Adler and Golany (2002) combined Principal Component Analysis with DEA to reduce the curse of dimensionality. PCA aims to find out the uncorrelated linear combinations of original inputs and outputs, and therefore to improve discrimination in DEA with minimal loss of information. Jenkins and Anderson (2003) used partial covariance analysis to identify those variables that could be omitted with least loss of information as measured by the proportion of total variance in all the variables lost by omitting particular variables.

In short, the number of DMUs should be at least three times the total number of inputs plus outputs used in the models. So, for example, if there are three inputs and five outputs, the minimum number of DMUs should be 24. This is a rule-of-thumb decision but from a practical point of view works reasonably well (Cooper et al. 2007). Often, another similar rule can offer guidance as follows: $n \geq \max\{m \times s,\ 3\times\ (m + s)\}$, where $n =$ number of DMUs, $m =$ number of inputs, and $s =$ number of outputs.

Too few variables, of course, reduce the model almost to a ratio measure – but this may still be usable for some Variable Returns to Scale type studies.

13.4.4 Relationships and Proxies

Data may be available on measures that really represent the same measure, although often expressed in different units. An example of this may be staff in a branch where data on both salaries and FTEs (Full Time Equivalent) are available. Of course, these represent the same variable, labor, so typically only one is used – but which one? The decision is dictated by the objectives of the study. More often, the FTE measure is used because this eliminates the dispute over pay scales that may be different depending on the local economic realities (large city vs. small community). However, if the manager has the flexibility of using staff in different capacities – less costly workers assisting a more costly one (really good sales people receive more support) the salary costs may be a better measure to bring out the efficiency gained by more effective management of the resource.

But there are other situations where different variables prove to be very highly correlated, even if they are not related logically. We may consider one as the proxy of the other. In this case, it is appropriate to use only one of the measures because the highly correlated other(s) only decrease the discriminatory power of the model without adding useful information. Hence, it is customary to run correlation analyses on all inputs and outputs selected for the model to see if one or more are highly correlated and then decide which may be dropped from the model. One might also take into consideration suspected relationships between certain input(s) and output(s) and test them to ensure that those that do correlate are left out of the analysis as appropriate.

A practical comment: it is always better to choose the variable that management sees as more representative of their view of the units' production model.

13.4.5 Outliers

There will always be problems with empirical data, either because some data elements are wrong or missing or because some DMUs are outliers and really do not belong to the dataset. Before meaningful DEA results can be obtained, these outliers must be dealt with.

The obvious first step is to double-check the data of those DMUs that appear to be performing too well, or too poorly. One way to find the former is to check the number of peers that use them as an efficient reference – it will be easy to see if this is much larger than for the other efficient DMUs. DMUs scoring very low, for example 0.2 or less, in a bank branch analysis are also suspicious. After all, in a typical bank, branches are closely controlled and have numerous policies, procedures, and rules of operation, so it is not possible to have branches that are 20% or less efficient in practice. Hence, checking data integrity is the first step, followed by either correcting data errors or removing problematic DMUs.

Removing outliers is justified only if they can be identified as having erroneous or missing data, or if, even when the data is correct, these DMUs really are in a different business than the others. For example, a "real-estate" branch will always have significantly higher loans and assets than other branches. It is, therefore, unrealistic or unfair to expect otherwise similar sized or located branches to emulate this activity.

There are several methods proposed for detecting outliers. Fox et al. (2004) derived sophisticated methods to detect outliers based on the defined "scale outlier," if it was relatively larger (or smaller) in all, or many, dimensions than other observations; and "mix outlier" if it was an unusual combination in terms of the size of vector elements relative to other firms. Cazals et al. (2002) and Simar (2003) proposed an order-m partial frontier to detect superefficient outliers. Based on the superefficiency DEA model proposed by Andersen and Petersen (1993), Avkiran (2006) identified the efficient outliers with superefficiency scores equal to 2 or above.

13.4.6 Zero or Blank?

When datasets are obtained from real operations, often we find a number of data items which are "blank." This is different from an entry that is actually zero because that means that the DMUs use or produce zero quantities of the input/output. A blank may mean zero, N/A (not available) or simply "we do not know." If there is no information or guidance from the data source about how to deal with these data items, the associated DMUs must be excluded from the analysis. If they are included, then there must be an agreement by everyone that whatever is used (e.g., zero) is acceptable.

13.4.7 Size Does Matter

At first glance, the size of the bank or branch appears to have significant effects on the operations of the unit. Surprisingly, while this is indeed the case for banking corporations, it does not hold for branches. Banks are best modeled using BCC models to allow for their size differences. Branches, on the contrary, tend to operate as CCR entities. Most banks first segment their branch networks into various groups based on business type and then evaluate them separately within their own group. Typically, large commercial branches are removed from the analysis altogether because different models would be used for them.

Retail branches are often segmented into four groups:

- Small Rural (small towns and villages)
- Small Urban (local residential areas in large towns and cities)
- Large Rural or Regional (located in larger towns and some branches serve local businesses as well)
- Major Urban (Large cities, sophisticated clientele, investment and business orientation).

This segmentation is often employed in DEA studies of larger banks where there are a sufficient number of branches.

Unfortunately, such distinctions are often the cause of inaccurate results because geographical location is not necessarily a good factor to decide on the similarity of their operating environments. Paradi et al. (2009) proposed a novel grouping approach in a DEA context and applied it to over 900 branches of a Canadian bank to identify branch managerial groups based on their operating patterns. The proposed DEA approach was compared with the bank's internal grouping method and a traditional clustering method to show the improvement in the choice of reference DMUs within groups.

13.4.8 Too Many DMUs on the Frontier

Banks tend to be very well managed, or at least well controlled, as they have to follow policies and rules laid out by the Head Office and the Government regulator. This results in a substantial portion of the branches being on the frontier, typically 25–50%. While this is not a problem with the technique per se, it is a problem if management wishes to improve operations across the branch network because frontier resident branch managers see themselves as already being the best they can be. Sowlati and Paradi (2004) addressed this issue by developing a management opinion based technique that created a "Practical Frontier" that enveloped the empirical one, thus offering targets to the empirically efficient units. Florens and Simar (2005) proposed a two-stage approach to generate a parametric approximation for nonparametric efficient frontier. The standard deviations of the parameters were obtained by a bootstrap algorithm.

13.4.9 Environmental Factors

In many real-world situations, DEA could be applied with considerable success, yet it is found that the typical DEA results are not satisfactory because management (the "measured") push back citing unfairness due to a lack of consideration of "external effects." Fairness and equitable treatment are both key components if people are to accept the outcomes from DEA (or any other) studies. These external factors can be classified as environmental factors. Examples of these may be the economic environment in one geographic area being different than that in another (a disadvantaged State/Province/District vs. an advantaged one), different opening hours, demographics, etc.

Another such factor, which is more difficult to quantify but is very important, is the competitive environment the bank or branch operates in. Clearly, if the branch is the only one in town, it gets almost all the business, but if it is located across the street from three other bank branches on the same intersection and has two or

three others within a block, it has to fight for its market share. An effective way to include this factor is to develop a "competitive index" as was done by Vance (2000). She incorporated in her index the type and number of competitors in the reasonable geographical drawing area of a branch.

13.4.10 Service Quality

Service quality is an important issue but it is difficult to measure and even more difficult to incorporate into the model. Typically, there are two subjective measures used when estimating such data: Customer Satisfaction and Employee Satisfaction. As data on both of these are acquired through questionnaires, quality control of the data collection process is a crucial issue. Often, the questions offer choices from a balanced five-step Likert scale response set. Banks tend to count, as the percentage of all responses, only the top two scores (Excellent, Very Good or if applicable the most negative two) and ignore the rest. This is not a very satisfactory process, but typically, it is the best available data.

13.4.11 Validating Results

For most managers, DEA is an unknown "black box" and without meaningful validation, they will not use it or believe in the recommendations offered on the basis of such results. Hence, validation of the results is important, although not at all easy. One method is to compare DEA results with the bank's own performance measures. Another technique is to use statistical methods to establish the credibility of the approach. But care must be taken with these because the need is to show how much better DEA is, and not leave the impression that it is no better than what they have already. The third approach is to conduct Monte Carlo experiments to examine the performance of the DEA model.

13.5 Banks as DMUS

Since 1985, DEA has been widely applied in the banking sector around the world. The top ten examined countries include: USA (50), Spain (28), Canada (25), UK (25), Sweden (23), India (22), Italy (20), Taiwan (19), Greece (18), and Germany (17). Most of cross-country studies focused on the European Union banking sectors (22). Among these studies, 36% of the DEA models are applied with the assumption of variable returns to scale (VRS), 26% with the assumption of constant returns to scale (CRS), and 38% of studies are conducted under both CRS and VRS assumptions.

13.5.1 Cross-Country/Region Comparisons

One of the characteristics of DEA is that it requires that all DMUs under examination be comparable and, therefore, have the same "cultural" background or that adjustments can be made for the differences between, for instance, banks in different countries/regions. Some studies that estimate bank efficiencies in a cross-national scenario simply build a common frontier and measure bank efficiency without considering their environmental conditions. This evaluation approach is biased by the differences in regulation, economic and demographic conditions, which are beyond the control of bank managers (Lozano-Vivas et al. (2002)).

Many applications use a traditional DEA model in the first stage to evaluate bank efficiency, and then in the second stage, use regression or other statistical techniques to examine the relationships between the DEA efficiency and various bank-specific, country-specific and macroeconomic factors, such as bank ownership, bank size, industry concentration, market capitalization, GDP growth, and the macroeconomic and legal environment (e.g., Casu and Molyneux (2003), Hauner (2005), Oliveira and Tabak (2005), Stavárek (2006), Pasiouras (2008), Delis and Papanikolaou (2009)). This approach fails to place all the DMUs on a level playing field or adjust efficiency measurements based on the environmental differences.

In the DEA banking literature, there are mainly two approaches that are used to generate adjusted efficiency measures according to the heterogeneous operating conditions. The first method is a single stage adjustment which incorporates the environmental effects directly into the model, such as including environmental variables as nondiscretionary inputs or outputs following the route provided by Banker and Morey (1986) (such as Lozano-Vivas et al. 2002). The brief mathematical formulation used in Lozano-Vivas et al. (2002) is presented as follows:

$$
\begin{aligned}
&\min_{\theta,\lambda} \theta \\
&\text{s.t.} \sum_{j=1}^{n} y_{rj}\lambda_j \geq y_{r0} \quad r = 1, \ldots, s \\
&\sum_{j=1}^{n} z_{kj}\lambda_j \geq z_{k0} \qquad k = 1, \ldots, q \\
&\sum_{j=1}^{n} x_{ij}\lambda_j \leq \theta x_{i0} \qquad i = 1, \ldots, m \\
&\sum_{j=1}^{n} \lambda_j = 1 \\
&\lambda_j \geq 0
\end{aligned}
\tag{13.1}
$$

where θ = the radial efficiency score; λ_j = optimal weights of referenced units for unit j; x_{ij} = value of the ith input to unit j; y_{rj} = value of the rth output from unit

Table 13.1 DEA Model for Lozano-Vivas et al. (2002)

Inputs	(1) Personnel expenses, (2) Noninterest expenses
Outputs	(1) Loans, (2) Deposits, (3) Other earning assets
Environmental variables	(1) Income per capita, (2) Salary per capita, (3) Population density, (4) Ratio of deposits per square kilometer, (5) Income per branch, (6) Deposit per branch, (7) Branches per capita, (8) No. of branches per square kilometer, (9) Equity over total assets, (10) Return over equity

j; z = the environmental outputs and any input is transformed into nondiscretionary output by reversing its sign and translating it; q = the number of nondiscretionary variables. Using this model, each bank is compared with a positive linear combination of banks for which the value of each environmental factor is not better than the corresponding value of the bank. The environmental features include macroeconomic and regulatory conditions as well as accessibility of banking services. The variables used in Lozano-Vivas et al. (2002) are listed in Table 13.1.

Lozano-Vivas et al. (2002) applied this model to investigate the operating efficiency differences of 612 commercial banks across 10 European countries, including 24 Belgian, 29 Danish, 150 French, 203 German, 26 Italian, 68 Luxemburgian, 22 Dutch, 17 Portuguese, 28 Spanish, and 45 British institutions. The comparison between the results obtained from the basic and the adjusted DEA model showed that the worse the country-specific environmental conditions, the greater the changes in the efficiency scores.

The second is multistage approaches to adjust the initial input/output dataset to reflect the effects of the operating environments (such as, Drake et al. 2006; Liu and Tone 2008; Avkiran 2009). The common process of a multistage approach is to, first, run a traditional DEA efficiency model, second, conduct regression analysis to quantify the effects of environmental factors based on the slacks obtained from the first stage model, third, adjust the original dataset, and finally, rerun the DEA model on the adjusted dataset to generate the final efficiency results. Liu and Tone (2008) applied a stochastic frontier analysis model to decompose the slacks, while Drake et al. (2006) and Avkiran (2009) employed Tobit regression models to account for the environmental effects.

13.5.2 Bank Mergers

One of the hot topics in the Financial Services Industry was the significant merger activity USA and European Banks engaged in late 1990s and into the 2000s. In USA alone, from 1990 to 2008, there were more than 100 merger and acquisition deals. Much work has been done on the efficiency effects arising from bank mergers using various methodologies, ranging from simple ratio analyses to more complex frontier approaches; both econometric and mathematical methods were used, such

Table 13.2 DEA studies relevant to bank mergers and acquisitions

Authors (year)	# of Banks	Country	Objectives
Resti (1998)	67	Italy	Estimate the relative performances of buyers, targets, and merged banks
Avkiran (1999)	19	Australia	Examine the role of mergers in bank efficiency gains during the deregulated period
Seiford and Zhu (1999)	55	USA	Examine the effects of acquisition on bank efficiency and attractiveness
Vela (2002)	2	Canada	Assess bank merger efficiency gains, estimate the synergy potentials from corporate culture adoption
Havrylchyk (2004)	9	Poland	Investigate the efficiency changes before and after bank merger
Sherman and Rupert (2006)	4	USA	Analyze merger benefits based the comparison of the branch operating efficiencies
Sufian and Majid (2007)	5	Singapore	Investigate the effect of mergers and acquisitions on the bank efficiency and its determinant factors
Al-Sharkas et al. (2008)	440	USA	Investigate the sources of efficiency gain associated with bank mergers
Wu (2008)	17	Australia	Investigate pre- and postmerger bank performance
Paradi et al. (2010a)	2	Canada	Identify the potential benefits that may result from sharing cultural advantages during the process of consolidation and restructuring associated with bank merger

as Berger (1998), DeYoung (1993), Peristiani (1993, 1997), Grabowski et al. (1995), Vander Vennet (1996), Resti (1998), Avkiran (1999), and Haynes and Thompson (1999). Some selected DEA studies relevant to bank mergers and acquisitions are listed in Table 13.2.

In the twenty-first century, the cross-border mergers and acquisitions era has emerged. Banks from many countries began to buy banks in other countries as illustrated in a partial list of these events, where €b = billion Euros and $b = billion US dollars (Table 13.3).

In addition to measuring postmerger gains, the importance of estimating the *potential* for achieving added benefits through the synergies arising from the adoption of a more advanced management style and operating infrastructure may be even more interesting. Here, the management style and operating infrastructure are defined as corporate "culture," and can be adopted from each of the merger partners by the merged bank for implementation in its combined branch network.

To illustrate the process, two banks of different cultural backgrounds wish to assess the potential impact on efficiency from merging their branches (Paradi et al. 2010a). One bank is positioned as a market leader in service quality, while

Table 13.3 Selected cross-border bank mergers and acquisitions

Year	Acquired bank	Acquired bank nation	Acquirer	Acquirer nation	Transaction value
2000	CCF	France	HSBC	UK	11.23(€b)
2000	Bank Austria	Austria	HVB Group	Germany	7.17 (€b)
2000	Unidanmark	Denmark	Nordic Baltic Holding	Sweden	4.78(€b)
2000	PaineWebber	USA	UBS AG	Switzerland	1.20 ($b)
2001	Banacci	Mexico	Citigroup	USA	6.26 ($b)
2001	Schroders plc	UK	Citigroup	USA	2.20 ($b)
2004	Abbey	UK	BSCH	Spain	13.85 (€b)
2005	HVB Group	Germany	UniCredit	Italy	15.37 (€b)
2005	Banca Antonveneta	Italy	ABN Amro	The Netherlands	5.87 (€b)
2005	Bank Austria Creditanstalt	Austria	UniCredit	Italy	2.70 (€b)
2007	Commerce Bancorp Inc.	USA	TD	Canada	8.50 ($b)
2007	Compass Bancshares	USA	Banco Bilbao Vizcaya Argentaria	Spain	9.80 ($b)
2009	Citicorp Diners Club Inc.	USA	BMO	Canada	1.00 ($b)

the other employs very effective corporate/investment strategies. Through simulating merger outcomes and taking the best approaches from the merger partners, a general approach to estimate possible efficiency gains can be developed in a step-by-step process.

13.5.2.1 Selecting Pairs of Branch Units for Merger Evaluation

This step involves the strategic selection of consolidating branch units based on intuitive criteria, such as branches in close physical proximity; similar business natures and objectives (i.e., distinguishing between general branches and processing cost centers), etc.

13.5.2.2 Defining a Strategy for Hypothetically Merging Two Bank Branches

Simulation of efficiency gains involves constructing a hypothetical unit that represents the combination of inputs and outputs of the two merged branches, less the expected reductions in costs specified by management. It is assumed that assets are not reoptimized for further cost savings, but rather staffing, back-office operations, and other inputs are rationalized to realize cost efficiencies. In measuring the potential output production of the combined entity, in the output levels a summation approach is used, since the business of the branch to be closed is assumed to be transferred to its merger partner, less some unavoidable attrition.

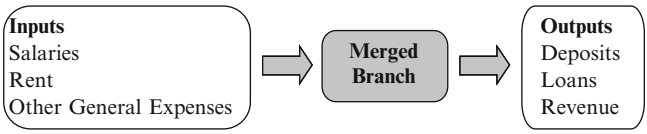


Fig. 13.1 Model for measuring efficiency gains from mergers

13.5.2.3 Developing Models for Evaluating the Overall Performance of Merged Units Through the Selection of Appropriate Input and Output Variables

An example of a model to evaluate the performance of a merged branch is depicted in Fig. 13.1. It includes salaries, rent, and other general expenses as inputs and three variables as outputs: deposits, loans, and revenues.

13.5.2.4 Calculating Potential Efficiency Gains

Performance gains are estimated by comparing the technical efficiency of the postmerger unit (hypothetically combined using the merging strategy defined above) to the corresponding premerger efficiencies of the merging branches (Vela 2002). More specifically, the change in technical efficiency (ΔTE) is calculated as the difference between the technical efficiency of the hypothetically merged unit (TE_M) and the weighted-sum of the premerger technical efficiencies of the two branches (TE_1 and TE_2) as shown in (13.2) below:

$$\Delta TE = TE_M - [w_1(TE_1) + w_2(TE_2)] \tag{13.2}$$

where w_1 and w_2 are the weights for the two branches before the merger, whose sum equals unity or $w_1 + w_2 = 1$. The weights are based on total assets (TA), such that $w_i = TA_i/(TA_1 + TA_2)$, where i represents Branch 1 or Branch 2 involved in the merger.

Efficiencies of the consolidating branches, as well as the hypothetically combined unit, are measured using standard DEA models. A sample of nonmerging units is also included in the analysis for comparison purposes.

13.5.2.5 Identifying Differences in Cultural Environments Between the Merging Banks

Differences in culture between the two banks are quantified using existing methodologies that provide relative index measures of cultural advantages or disadvantages. For the given example, the culture represents corporate strategies (corporate index or CI), along with organizational core competencies

that affect service quality (service index or SI). The cultural framework can be represented mathematically as follows, where *F* indicates *as a function of*:

- Corporate Strategy = *F* (Resources, Cost of Input, Product Portfolio, Price of Output)
- Service Quality = *F* (Human Capital, Technology, Operating Processes, Size of Activity)
- Overall Performance = *F* (Corporate Strategy, Service Quality, Branch Size, Cost, Products, Revenue)

Over time, the two branches merged into a single entity are assumed to adopt each other's more favorable or stronger cultural elements, i.e., more efficient investment strategies or higher level of customer service. This cultural adoption is reflected in adjustments of the level of inputs consumed and/or outputs produced.

13.5.2.6 Calculating Potential Synergies

A merged organization is assumed to eventually increase the level of revenues generated with the implementation of the corporate strategies from the culturally favorable branch. The overall increase is calculated by multiplying the revenues of the branch employing less efficient corporate strategies, i.e., relatively unfavorable, with the ratio difference of CIs associated with favorable (F) and unfavorable (UF) units, i.e., CI_F/CI_{UF}. Also, the merged branch is assumed to produce higher levels of customer service, by embracing the methods of the favorable merger partner. Similarly, increases are calculated using the ratio difference of SIs associated with favorable and unfavorable units, i.e., SI_F/SI_{UF}. The final values for the merged branch's inputs, deposits, and loans, along with revenues adjusted for culture (X_C, $Y_{D\&L,C}$, and $Y_{R,C}$, respectively), are calculated as follows (Vela 2002):

$$\begin{aligned}&\left(X_C, Y_{D\&L,C}, Y_{R,C}\right)\\&= \left[X_F + X_{UF}\left(\frac{SI_F}{SI_{UF}}\right), Y_{D\&L,F} + Y_{D\&L,UF}\left(\frac{SI_F}{SI_{UF}}\right), Y_{D\&L,F} + Y_{D\&L,UF}\left(\frac{CI_F}{CI_{UF}}\right)\right]\end{aligned} \tag{13.3}$$

where X_F and X_{UF} are the original input values of the branches that operate in favorable and unfavorable cultural environments, respectively. The same notation is used for Y_U and Y_{UF}.

Potential synergies are estimated as the difference in efficiencies between the hypothetically merged branch and the same branch that has undergone cultural adoption, which is also calculated using standard DEA models. The final step involves testing whether additional gains in efficiency are possible when the impact of cultural environments is incorporated into the analysis.

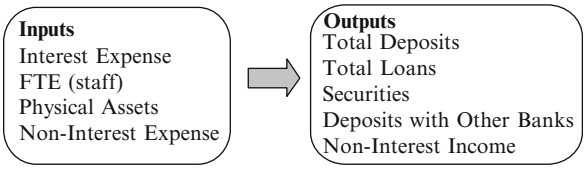


Fig. 13.2 Variables in the production model

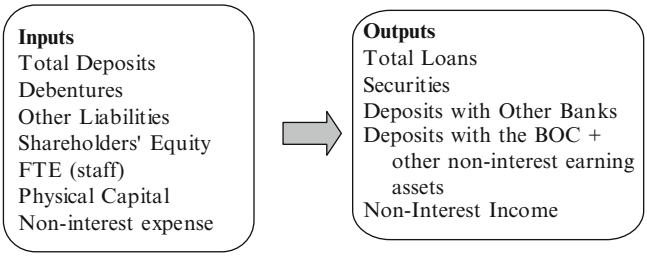


Fig. 13.3 Variables in the intermediation model (Aggarwall 1996)

13.5.3 Temporal Studies

In many instances, progress over time becomes the main objective for measuring the efficacy of management strategies. Windows Analysis and Malmquist Index are two interesting techniques that one can employ to deal with time series data and gain valuable insights from the pattern and direction of changes during the study period.

13.5.3.1 The Models

Most of DEA studies focus on the technical efficiency of banks (Fethi and Pasiouras 2010). The Production model illustrated in Fig. 13.2 can be thought of as showing how well the banks perform their production of tasks, deal with customers and make transactions as required (Aggarwall 1996).

Other noninterest earning assets such as accrued interest, deferred and recoverable income taxes, goodwill, accounts receivables and prepayments, deferred items and some other assets are not considered in the model. It is neither a management objective to increase these assets, nor do they qualify as a service provided by the banks. There is also no interest earned on this category of assets. The results from this model are referred to as the Technical Efficiency results.

.

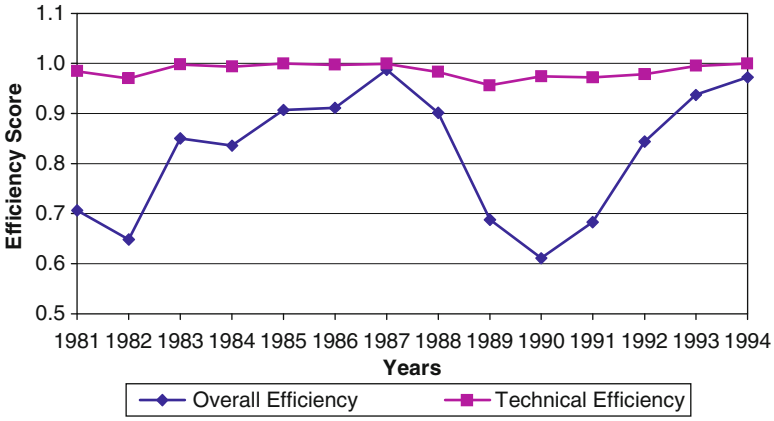


Fig. 13.4 Technical and overall efficiency for all banks for each year

The model illustrated in Fig. 13.3 can be considered as the model that explains how well the banks does in their efforts to make a profit by dealing with money and securities in a risky environment. The results from this model are referred to as the Overall Efficiency results in Fig. 13.4.

The weights in both models are constrained using real prices and management provided information as appropriate (Aggarwall 1996). But again, these developments are not shown here.

Table 13.4 gives the statistics on the data used.

13.5.3.2 Window Analysis

One of the challenges in DEA modeling is the situation when there is an insufficient number of DMUs in comparison to the number of relevant inputs and outputs in the model. One of the approaches is to collect a time series panel data and then use the DEA Window Analysis approach. This technique works on the principle of *moving averages* (Cooper et al. 2007) and is useful to detect performance trends of a unit over time. Each unit in a different year is treated as if it were a "different" unit. In doing so, the performance of a unit in a particular year is compared with its performance in other periods in addition to the performance of the other units. This means that the units of the same DMU in different years are treated as if they were independent of each other – but comparable.

A notable feature of this technique is that there are nk units (DMUs) in each window where n is the number of units in a given time period (and it is the same in all time periods), and k is the width of each window (equal for all windows).

Table 13.4 Data statistics

	Min	Max	Average
Inputs[a]			
Interest expense	1,790,898	16,358,317	6,391,627
FTE (staff)[a]	10,877	52,745	30,280
Physical capital	124,941	2,057,000	862,912
Other noninterest expenses	216,881	1,837,556	779,608
Total deposits[b]	23,166,413	135,815,000	78,237,069
Debentures	25,157	282,182	133,098
Other liabilities	996,507	27,748,000	5,664,598
Shareholders' equity	825,683	8,589,000	4,419,406
Outputs[a]			
Total deposits[b]	23,166,413	135,815,000	78,237,069
Total loans	19,626,193	117,013,807	65,867,548
Securities	2,202,023	28,753,000	10,781,395
Deposits with other banks	630,112	22,345,822	8,396,201
Deposit with BOC + other noninterest earning assets	762,946	8,791,198	3,707,324
Noninterest income	195,071	2,697,000	961,179

[a] All figures except FTE are in thousands of Canadian Dollars
[b] Total deposits appear as input in one model and as output in the other

This feature is extremely important in the case of a small number of DMUs and a large number of inputs and outputs since it increases the discriminatory power of the DEA models. This is accomplished by dividing the total number of time periods, T, into a series of overlapping periods or *windows*, each of width k ($k < T$) and thus having nk units. Hence the first window has nk DMUs for the time periods $\{1, \ldots, k\}$, the second one has nk DMUs and the time periods $\{2, \ldots, k + 1\}$, and so on and the last window consists of nk DMUs and the time periods $\{T - k + 1, \ldots, T\}$. In all, there are $T - k + 1$ separate analyses where each analysis examines nk DMUs.

An important factor is the determination of the window size. If the window is too narrow, there may not be enough DMUs in the analysis and thus not enough discrimination in the results (which also depends on the number of DMUs and variables in the model). On the contrary, too wide a window may give misleading results because of significant changes that occur over periods covered by each window. The best window size is usually determined by experimentation.

Panel data, from the six largest Canadian banks during a 14 year period (1981–1994) was used in the analysis (Aggarwall 1996). However, during the 14 year period a substantial difference can be expected in the production possibilities of the banks due to changes in the laws governing banks, acquisitions, technology growth and even leadership changes. Therefore, it is not reasonable to compare DMUs over such a long and disparate period in one analysis. Hence, the window analysis technique is used to obtain efficiency results for a few (just six DMUs) banks over time and to identify performance trends.

The first step of a Window Analysis is to select a *window size*. For this model (with output multipliers constraints), not much interpretation (discrimination) is possible

using 3-year windows ($k = 3, n = 6$). This is due to the insufficient number of units in each window (only 18) compared to the number of variables (nine). Four-year windows are not discriminatory enough either; hence, *5-year windows* are selected as the narrowest window that still enables a reasonable discrimination for the analysis.

Using the DEA model described above with 5-year windows, it is evident that the efficiency of the banks reflects the economic and managerial activities that took place during the time period analyzed. Table 13.5 shows the results for one of the banks. From all the tables such as these a comparison of the various bank performances can be made.

The results of both models are presented here and the Overall Efficiency results only are shown (not how they are produced). When these results are combined with the Technical Efficiency detailed above, a graph is constructed for all six banks as a group to show the industry trends as can be seen in Fig. 13.4.

The striking observation here is how well the overall efficiency reflects the economic events of the 1980s and 1990s in Canada. A boom through the latter part of the 1980s followed the recession in the early part of the decade. The performance again declines with the next economic down cycle as the 1990s arrive and then rises as expected when the economy rebounded later in the decade.

Technical efficiency, on the contrary, shows a very consistent record as the banks are in a lock step with each other in a very highly controlled and extremely competitive environment where producing transactions is not effected much by the economic environment.

This section illustrates a way in which a small number of DMUs can be analyzed when a temporal database is available. There are more applications of DEA windows analysis in the banking sector, such as Webb (2003) investigated the relative efficiency levels of seven large UK retail banks during the period of 1982–1995; Asmild et al. (2004) evaluated five Canadian banks' performance and productivity changes over 1981–2000; Sufian and Majid (2007)) measured the trends in the efficiency of six Singapore banks over 1993–2003.

13.5.3.3 Malmquist Productivity Index

The banking industry is continually being shaped by market forces, which include worldwide competition, technological advances in production processes and nonbanking firms competing for business. The continuous disintermediation is being accelerated by technology that enables competitors to enter markets merely by using the Internet and the World Wide Web. Domestic banks in countries that traditionally were almost oligopolies also face erosion of their market share by new competitors, such as "white label" ATM machines, large retailers that provide a "cash back" service to customers, foreign banks doing business strictly electronically, and via the Internet. Hence, productivity gains for banks will partially be caused by technological advancements and partially by more efficient management. How does one separate these effects? One method is the Malmquist index (for the technical development of this approach see the book by Coelli et al. (1998) and Cooper et al. (2007)).

Table 13.5 Technical efficiency scores from the output constrained model with 5-year windows for one bank

Window	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
1981–1985	0.638	0.603	0.898	0.918	1									
1982–1986		0.553	0.830	0.848	0.931	1								
1983–1987			0.702	0.715	0.777	0.829	1							
1984–1988				0.715	0.777	0.829	1	0.951						
1985–1989					0.777	0.829	1	0.951	0.771					
1986–1990						0.829	1	0.951	0.771	0.779				
1987–1991							1	0.951	0.771	0.779	0.894			
1988–1992								0.773	0.606	0.607	0.706	1		
1989–1993									0.526	0.521	0.610	0.870	0.988	
1990–1994										0.493	0.579	0.829	0.928	0.993
Average	0.638	0.578	0.810	0.799	0.852	0.863	1.000	0.915	0.689	0.636	0.697	0.900	0.958	0.993

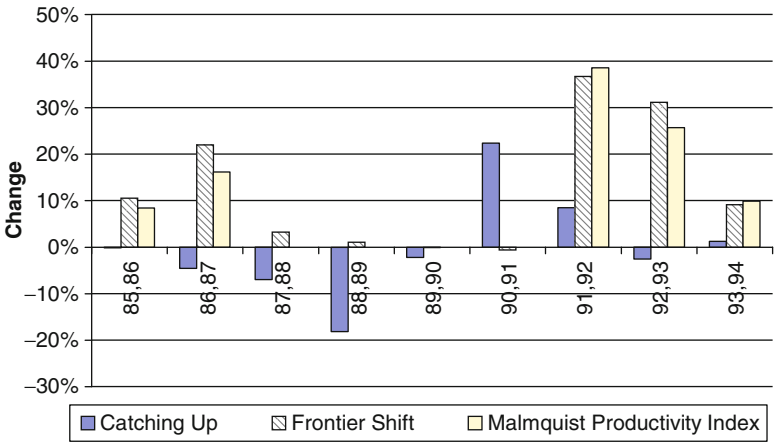


Fig. 13.5 Malmquist productivity index components for adjacent periods

Productivity growth measurement involves separating changes in *pure productivity* (level of inputs necessary to produce a level of outputs) from changes in the *relative efficiency* of the DMUs over time. The Malmquist Productivity Index can be used to measure the *productivity growth* of DMUs between any two time periods t_1 and t_2. In this method, the relative distances of the DMUs from the frontier are used in conjunction with the relative changes in the position of the frontier from one period to the next to reveal the total changes in the productivity of a DMU.

In the Malmquist Productivity Index approach, a DMU is studied at two different points in time with respect to a common reference technology. Hence, *productivity* change is measured as the ratio of distances to the common technology frontier, while *technical efficiency* change can be expressed as the relative distance from an observation to the frontier, keeping constant observed proportions between outputs and inputs. The path to the frontier can be accomplished either by input reduction or by output augmentation. The Malmquist Index can then be expressed as a ratio of corresponding efficiency measures for the two observations of the DMU.

The mathematical formulation of the Malmquist Productivity Index can also be found in a number of papers (c.f. Malmquist (1953), Caves et al. (1982), Berg et al. (1992), Färe et al. (1994), Ray (2004), Cooper et al. (2007), Asmild and Tam (2007)). In this section, the results from the 14-year panel data (Aggarwall 1996) are shown in Fig. 13.5 applying the Malmquist Productivity Index to DEA window analysis scores. The figure shows the catching up, frontier shift, and Malmquist Productivity Index change for the average of the six banks studied. The technology is fixed for 1985.

As can be seen from Fig. 13.5, the banks, on average, were not keeping pace with the frontier's movement as their catching up component is negative for five out of the nine periods examined here. When the changes are evaluated over longer periods of time (1985–1994, 1985–1989 and 1989–1994) a similar picture emerges, as seen in Fig. 13.6.

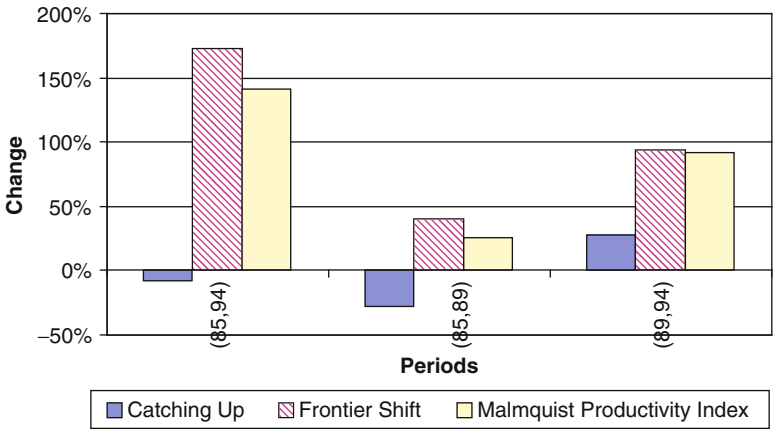


Fig. 13.6 Average productivity change for the banks over longer periods

These results indicate that the Canadian banks increased their productivity quite significantly during the 10-year period presented here; in fact, they more than doubled the services delivered per unit resource consumed from 1985 to 1994. However, it is evident that such growth is mainly due to the frontier shift that resulted from the massive technology deployment that occurred during the period. The catching up component is actually negative for the first 4 years and for the entire 10-year period as shown by the bars in Fig. 13.6. The conclusion from this is that the banks have actually been falling behind in management induced productivity rather than catching up. The lesson for management is that there are opportunities to improve productivity in the catching up dimension and that this could make a significant contribution to increased efficiency.

Asmild and Tam (2007) proposed a concept of "global" Malmquist indices and "global" frontier shift indices, which were based on the calculation of the distance for all DMU's relative to each time period's frontier. They further applied this method on a sample of 115 branches of the same bank located in six different countries to estimate the differences between the domestic frontiers.

Recently, many studies have applied the Malmquist Productivity Index to analyze the effects of legal and financial events on bank productivity. Mukherjee et al. (2001) explored productivity growth for a group of 201 large US commercial banks over the initial postderegulation period from 1984 to 1990. Sathye (2002) assessed the effects of deregulation and the reforms on the productivity over the period of 1995–1999 using a panel of 17 banks. Al-Muharrami (2007) examined the productivity development in 52 Arab GCC banks during the period 1993–2002. Tortosa-Ausina et al. (2008) explored the productivity growth for 50 Spanish savings banks over the postderegulation period 1992–1998. Liu (2010) employed the Malmquist productivity index approach to measure the technical efficiency and productivity change of the 25 commercial banks in Taiwan over the post-Asian crisis period 1997–2001.

13.6 Bank Branches as DMUS

From many aspects, bank branch studies are more desirable because from a managerial point of view, improvements can be achieved directly by the branch managers and so results from the analyses affecting the bottom line are close at hand. In cross-bank analysis the identified benchmarks may be hard to use as role models because the difficulties in relating to what others are doing gives rise to criticisms of unfair and inequitable outcomes based on what demands are made on people. Therefore, there are a larger number of practical analyses where bank branches of the same bank were used as DMUs.

Traditional productivity measures were first devised in manufacturing concerns where absolute standards were available for comparison. All participants knew what a "good job" was and the standard was not debatable. Moreover, when productivity fell, the reasons were relatively easy to find and remedied. But services industries are fundamentally different organizations and while the inputs are no more difficult to identify (i.e., labor and other operational expenses), the outputs are much more troublesome to properly define and more importantly, measure. Some of the outputs may be hard to measure such as the following: how well is the customer served? How many transactions per period are possible per staff member? What types of transactions are appropriate to measure, each with varying degrees of difficulty and time? What are the metrics for these different types of transactions? Moreover, there are no theoretical maximums for these measures. Hence, reaching 100% productivity cannot be observed either.

To illustrate the methodology of how to build bank branch models, data from a large Canadian bank's branch network were used in the study of Paradi et al. 2010b. The dataset had 816 branches with very extensive and detailed data items for each branch.

The objective of this example is to evaluate branch performance along three dimensions: Production, Profitability, and Intermediation. Three DEA models were developed to identify the improvement opportunities in each dimension. This approach improves the reality of the performance assessment method and enables branch managers to clearly identify the strengths and weaknesses in their operations. Psychologically, when someone sees that he/she excels in his/her efforts in some way, he/she is much more likely to accept suggestions on how to improve in other areas – especially when suggestions on learning from their peers are included. In other words, by being perceived as fair and equitable, any measuring tool or system has a much better chance to be accepted and acted upon than when the opposite is true.

In this example, both CCR and BCC models were applied to investigate the scale effects on a branch's performance. Based on the review of 68 DEA studies in bank branch analyses from 1997 to 2010, it was found that 46% use CCR assumption, 22% use BCC assumption, and 32% use both CCR and BCC assumptions.

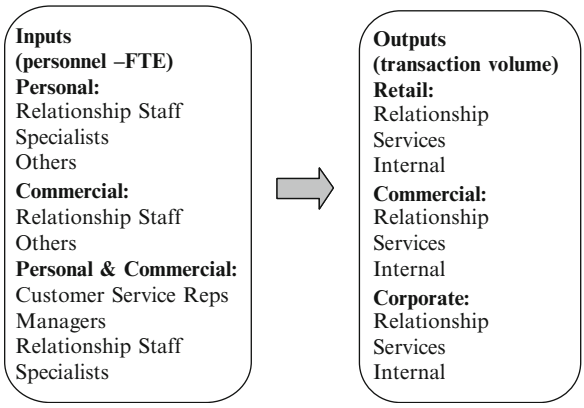


Fig. 13.7 Production model inputs and outputs

13.6.1 The Production Model

In bank branch analyses, the production model commonly views bank branches as producers of services and products using labor and other resources as inputs and providing deposits, loans and others (in value or number of transactions) as outputs. The transactions may be face-to-face with the customer in the branch, carried out in the back office or even delivered at customer premises. The branch is a service "factory" and customer satisfaction is also a key outcome of a good effort. The model shown in Fig. 13.7 is designed to measure staff performance in a branch in terms of producing transactions (Paradi et al. 2010b).

In the example of Paradi et al. 2010b, on the output side, the transactions were separated according to customer types and the difficulty required to complete the transaction. The examined Bank suggested three main customer types: Corporate (large business, i.e., Wal-Mart, Microsoft, and Air Canada), Commercial (small- and medium-sized businesses), and Retail (individuals). The transactions under each customer type were further split based on the transaction difficulty. The relationship transaction included mortgage and loan applications and approvals, as well as retirement plan transactions. Service transactions included deposits, withdrawals, money orders, and account inquiries. Internal meant back-office transactions with little or no customer contact including error corrections, accounting entries, charge backs, ABM servicing, and others. In totally, there are nine different types of transactions on the output side.

On the input side, staffing is the most important branch operating expense, often accounting for up to 75% of the total. Combining all different types of staff together may lead to a confusing result due to their different responsibilities (and salary costs). In the above model, the staff was categorized according to business lines: Personal, Commercial, and Personal and Commercial. Under each business line, the staff was further classified according to different services performed. In total, nine staff types were identified on the input side. Personal and Commercial Specialists handle mortgages, loans, and the like; Personal Specialists handle investment

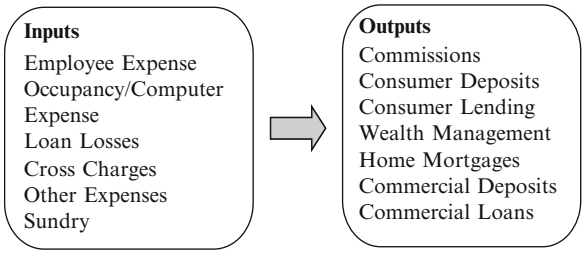


Fig. 13.8 Profitability model inputs and outputs

planning, private wealth management, and retirement planning; Customer Service Representatives (CSRs) handle face-to-face customer service transactions; relationship personnel handle a broader range of CSR duties including lending; managers are branch supervisors and managers; and other personnel handle a variety of transactions including accounting and back office.

13.6.2 Profitability Model

The profitability model is designed to examine the process of how well a branch uses its inputs (expenses) to produce revenues (Paradi et al. 2010b). This is fairly straightforward because it treats the branch as the producer of a product as opposed to the offerer of a service. While this does simplify the identification of inputs and outputs there are still some complexities to address. There is the issue of separating some revenues from their products. For example, a bank provides a below prime interest loan to a customer but requires that a certain percentage of the funds lent be held in the bank account (which pays minimal or no interest to the customer). This results in certain products appearing more profitable because the customer is paying interest on money that the bank has not released in practice. In this case, there is a lower lending revenue (because of the below prime interest rate) but higher commercial banking revenue from the interest earned on the portion left on deposit at the bank. The model then looks at those measures that show how well the branch is doing in producing profits as seen in Fig. 13.8.

On the input side, expenses included were those that branch management was able to directly influence; others such as depreciation and capital expenditures were excluded. Sundry were negative revenue charges incurred by the branch. Loan losses, while clearly an output measure, were included as an input because the need was for minimizing this outcome, thus penalizing those branches with higher losses.

On the output side, variables used were revenues from all of the branch's lines of business including commissions generated, interest income, and noninterest income. It should be noted that if a branch is inefficient when analyzed using this model, we may need to further investigate the underlying reasons when relevant information is available. This is needed because since some investment decisions that determine revenues are made at the bank level, which is beyond the branch managers' control.

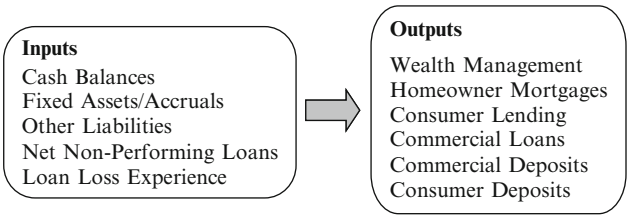


Fig. 13.9 Intermediary model inputs and outputs

13.6.3 Intermediation Model

The branch's intermediary role is mainly studied to examine how efficient the branch is in collecting deposits and other funds from customers (inputs) and then lending the money in various forms of loans, mortgages, and other assets (i.e., investments, etc.). A branch's intermediation efficiency is a strong indicator of the strength of its lending ability, which is, in turn, directly tied to a bank's ability to operate as a going concern. This approach was used in some of the earlier banking studies, such as Colwell and Davis (1992). Deposits, liabilities, and assets are the "raw materials of investible funds" (Berger and Humphrey 1990).

As mentioned earlier, there has been much debate on whether to include deposits as an input or output. Characterizing deposits as an input unfairly penalizes branches for seeking new depositors – an important sales function – and banks generate a significant amount of noninterest revenue from handling deposits. However, deposits are consistently positioned as an output under the user cost approach (Fixler and Zieschang 1999). Colwell and Davis (1992) also found that using only assets (loans plus investments) that earned interest, but since this model excluded other assets, they inflated the unit costs of larger banks. Their model used deposits as an input to illustrate this alternative.

Loan quality and losses are critical factors in a bank's health. The majority of research on the causes of bank failures finds that there is a strong relationship between the proportions of nonperforming loans and bank failures (Berger and Deyoung 1997). Hence, loan losses are usually included in the intermediation model, but on the input side, to appropriately reward those with small loan losses. On the contrary, too low loan losses may also be a problem because that implies that the bank is also passing up good business by being overly restrictive in its credit criteria.

Figure 13.9 shows an example of the selection of measures and the placement into input or output. The assumption is that the branches, to maximize income, should attempt to lend or invest as much money as possible, thus minimizing the funds on hand.

Table 13.6 Model orientation possibilities

Model	CCR-I	BCC-I	CCR-O	BCC-O
Production	X	X		
Profitability	X	X	O	O
Intermediation	X	X	O	O

13.6.4 Model Results

It is easy to see that a considerable effort must be made to determine which variables to include and where (Paradi et al. 2010b). The research purpose and context are essential in choosing inputs and outputs. Frequent management consultation is a must to ensure that all items are considered and approved by those who will eventually have to execute the suggested changes. Buy-in is essential for successful DEA result implementation. The above three models, all constructed during the same study and using the same data source, illustrate an approach that is likely to provide considerable amounts of useful information to the branch manager while being regarded as both fair and equitable.

From a practical point of view, branch managers can see what aspects of their branch need improvement and where they are doing well. It is not at all unusual for a branch to show efficient performance in one or two models while they could improve in the third. Moreover, they can contact other branch managers with whom they can exchange information on how to improve, and often this will be a mutual exchange of information in their respective areas of excellence. A spillover effect of this process is the morale level, which should not be negatively effected since they are exchanging information on a peer-to-peer basis, and neither has to feel inferior to the other. It is not a "competition" after all.

Most models can be processed either as input oriented or output oriented and sometimes both are carried out for the same model. Table 13.6 shows the possibilities in this set of three models. The "X" represents the orientation and technology actually used in this work (Paradi et al. 2010b) and the "O" represents orientation and technology that could have been used but was not. Only the Production Model is not a typical candidate to be used in an output-oriented analysis because the branch cannot do much, if anything, about getting customers into the branch to do business.

Table 13.7 shows the results from different models for the 816 branches of this large bank. This is a typical situation where 25–40% of the branches are DEA efficient, whatever the model. Understandably, this makes it difficult for senior managers to motivate those who are already on the frontier. The advantage of using three different models is that it is seldom the case that the same branch is efficient in all the models. Hence, opportunities for improvement will be present.

However, this does not permit us to examine the individual model's effects on the whole system and the trade-offs involved in focusing on certain areas. Managers may choose to focus on only one or two aspects of their branch as part

Table 13.7 DEA results summary

		Production model		Profitability model		Intermediation model	
Model	Efficiency scores	BCC	CCR	BCC	CCR	BCC	CCR
Global	% Efficient	33%	21%	38%	26%	29%	20%
	Average	0.77	0.71	0.87	0.82	0.81	0.76
	Minimum	0.28	0.25	0.32	0.26	0.34	0.29
	SD	0.21	0.20	0.15	0.16	0.17	0.18

of a competitive or purely reactive strategy based on their specific operating environments. A branch attempting to improve its lending results (or intermediation efficiency) could potentially lower their production efficiency. For example, to improve their average loan quality and reduce loan defaults, a branch may choose to increase the number of loan officers and the amount of time they spend with each client. However, this may well have a negative impact on the branch's production efficiency, as there is an increase in staffing (additional loan officers) combined with a decrease in the number of transactions performed (due to more time being spent with each client). Additionally, there could also be adverse effects on profitability (due to the increased staffing costs) if there was not a subsequent reduction in loan losses to offset the cost increase.

13.6.5 Senior Management Concerns

Unfortunately, most DEA studies of bank branches limit themselves to only one point of view, typically that of the individual manager or executive they work with or for. However, banks tend to be large and complex organizations in a managerial sense. Hence, models, even as focused as those shown thus far, might not meet the needs of the banks' top managers when you consider that they have the task of managing progress in all aspects of the firm. Therefore, they tend to use some ranking mechanism that will sort their branches from best to worst using either a single ratio, or most often a combination of "important" ratios. This approach results in rankings from 1 to n but the branches in the last half of the rankings may not be properly motivated to improve themselves in the standings. After all, what difference does it make whether you are 659th or 597th?

Clearly, DEA can be a very useful tool to offer managers valid and achievable goals. However, the typical DEA study shows a substantial number of branches as efficient and that often means that there is little differentiation among too many branches. Therefore, some senior managers reject DEA as a valid process for their firms when they seek continued performance improvements. In any case, senior management needs tools appropriate to their management processes.

Table 13.8 Model results

	SBM additive model
Average	0.78
SD	0.13
Median	0.78
Min.	0.37
% Efficient	7%

13.6.6 A Two-Stage Process

A ranking scheme enables senior management to target branches in most need of assistance and to reward those branch managers who perform exceptionally well. A way to address top management needs is to conduct a two-stage study (Paradi et al. 2010b). For example, one can carry out the three separate studies shown above and then devise a second stage to combine results of the three models and generate a comprehensive ranking index together with targets for improvement available. The application of a scheme that encompasses different performance measures clearly improves branch ranking accuracy.

This second-stage (or combined) model is formed by using the results from the initial three models as outputs. On the input side, a dummy variable with the value of 1 can be employed for all branches with the assumption that the bank fairly and equally supports all branches for providing financial products and generating profits.

In the study of Paradi et al. 2010a, b Slacks-based measure (SBM) Additive model was used to provide a more representative measure of efficiency. This resulted in efficiency scores being derived that accounted for all inefficiency sources. The SBM Additive model can be stated as shown in (13.3) from Cooper et al. (2007), where s_r^+ = output slack/shortfall for the r^{th} output; s_i^- = input slack/excess for the i^{th} input.

$$\rho = \left(\frac{1}{m}\sum_{i=1}^{m}\frac{x_{i0} - s_i^-}{x_{i0}}\right)\left(\frac{1}{s}\sum_{r=1}^{s}\frac{y_{r0} + s_r^+}{y_{r0}}\right)^{-1} \tag{13.3}$$

As the input was a dummy variable of 1 this eliminated the first half of the equation (the x_{i0} terms are all 1 and the input slacks, s_i^-, are all 0). A DMU is considered efficient if $\rho = 1$; this condition is met only when all slacks (s^- and s^+) are equal to zero (see Cooper et al. 2007). Owing to the elimination of the first half of the equation, the SBM Additive model is simply the average of the inverted scores. Using these scores, as opposed to just the average of the three scores, makes sense. The inverted scores penalize branches for being *overly* bad in one area. For example, a branch with inverted output scores of 0.9, 0.9, and 0.9 would have a combined score of 1.11 (or 90% efficient). Conversely, another branch with scores of 1.0, 1.0, and 0.7 would have a combined score of 1.14 (or 88% efficient). The simple mean of both branches is the same (0.90) but because the inverses are

Table 13.9 Connections for work being done

Staff type/transactions	Teller	Ledger handler	Accounting	Supervision	Typing	Credit
Business						
Counter interaction	X	X		X		
Counter sales		X	X	X		
Securities			X	X		
Deposit sales			X	X		
Commercial loans			X	X	X	
Personal loans			X	X	X	
Maintenance						
Term			X	X	X	
Commercial loan				X	X	X
Personal loan				X	X	X

used, the branch with the 0.7 output score is penalized for performing poorly in one area. Table 13.8 presents the statistical descriptions of the results of the 816 branches obtained from the SBM model.

As can be seen, the second-stage SBM model yielded significant benefits for senior managers over any single stage model, even if there were three as shown here. First, the model reduced the portion of efficient units from a range of 29–38% for the BCC model and 20–26% for the CCR model to just 7%. Managers can gain substantial new knowledge about their organization's overall performance and, if they implement the DEA suggestions, improve in a meaningful way.

13.6.7 Targeted Analysis

With all the tools available to the DEA analysts today, targeted work can be planned and executed with excellent results. The following is an example where management was planning to change the way they operated the back office in their branch work. The idea was to concentrate the back-office work to a number of regional centers where specialized processes with trained staff could do the work more efficiently than in the branch. If this worked, they could either reduce the workforce in the branch or free them up for more person-to-person customer work. The question was this: should this be done?

A dataset of 291 branches of a large Canadian bank was used for the study (see Schaffnit et al. 1997). A personnel model was constructed with management's assistance. Six types of employees were identified and six business transactions (involving customer contact) plus three maintenance or back-office transactions (not involving direct customer contact) were agreed upon. The transactions were used as outputs while the staffs (counted as FTEs) were the inputs. The interactions between staff and services performed are shown in Table 13.9. Notice, however, that not all staff were able (trained) to perform all required transactions.

Table 13.10 Inputs and outputs used in literature

Paper	Inputs	Outputs
Cook et al. (2000)	Service FTE	Deposits
	Sales FTE	Transfers between accounts
	Support FTE	Retirement savings plan openings
	Other FTE	Mortgage accounts opened
Cook et al. (2004)	FTE staff	Service transactions
	Operating expenses	Sales transactions
Portela and Thanassoulis (2007)	No. of electronic teller machines	Internet new registrations
	Rent	Transactions in cheque dispenser machines
	Not registered clients	Deposits in ETMs

Various input-oriented models can be applied here, but performance can only be fairly measured if appropriate constraints are placed on the multipliers. A BCC input oriented model was chosen and applied first to the whole model and then to a reduced model where the last 3 outputs were omitted.

The findings indicated that if the bank removed all the back-office work from all the branches, many branches with minimum staff configuration would end up with reduced workloads but without an opportunity to do something else. So if the bank did implement these plans across the branch network, it would result in a significant productivity loss as some people would have had nothing to do once the back-office work was removed from the branch.

13.6.8 New Role of Bank Branch Analysis

Recent development of alternative service distribution channels for bank branches, such as telephone banking, Internet banking, and automatic banking, increased the opportunities for banks to reshape their branching strategies, which may allow them to devote effort to more value-added services. There are several studies investigating this new function of branch branches using DEA models. Cook et al. (2000) developed a DEA model to separate an aggregated efficiency score into the sales and service component. They applied this methodology to a sample of 1,300 branches from a Canadian bank. Cook et al. (2004) studied the impacts of the introduction of electronic branches on productivity gains. Portela and Thanassoulis (2007) assessed branch performance in fostering the use of new transaction channels, such as automatic teller machines and cheque dispenser machines. Table 13.10 lists the inputs and outputs used in these three studies.

Table 13.11 Overall results by region (Paradi and Schaffnit 2003)

Region	D1	D2	D3	D4	D5	D6	D7	D8
Growth factor	2	2.5	3.3	3.3	2.7	2.5	2.9	3.3
# Branches	10	12	27	13	6	2	8	12
# Effective	7	1	1	1	4	0	1	2
% Effective	70	8	4	8	67	0	12	17
Average score	0.91	0.4	0.38	0.36	0.83	0.36	0.62	0.6

13.6.9 Environmental Effects

To generate reliable and acceptable branch efficiency estimations, the effects of "external" factors, which can affect a branch's operation processes but are beyond the branch manager's control, should be adjusted. In the published DEA applications, there are mainly two kind of environmental factors from a branch point of view: region specific and corporate-specific. When one compares branches across banks, differences in the corporate culture are inherent, as top management tends to determine what segment of the banking business they want to focus on. Hence, some methodology must be introduced to allow for the systemic differences caused by the managerial direction the branches are getting. Regional characteristics include economic growth rate, local community types, unemployment rate, etc.

Much work has been done to address these concerns and the following is an example where economic conditions are incorporated in the models along with a measure of the branch's business risks in its portfolio of loans. The dataset is of 90 commercial branches of a large Canadian bank, across eight economic districts in Canada (Paradi and Schaffnit 2004).

Several models were built and both input- and output-oriented BCC models were utilized. The regional economic data was obtained from the Bank's economics department and represented the average rate of change of the real regional domestic product. This factor was used to avoid unfair comparisons between branches operating, for example, in the economically disadvantaged Atlantic region of Canada to their West Coast counterparts, where the economic conditions were a lot more favorable.

The Table 13.11 shows the results of the study, district by district, restricted by the environmental variable – regional domestic product (growth factor).

Clearly, Region D1 has the highest percentage of effective branches. However, this is the undesirable result of using the economic constraints on peer selection because this Region, having the lowest economic growth rate, selects peers only from among others in the Region. Hence, unrestricted analysis must also be considered before final managerial decisions are taken. This is a good illustration of the problem one can encounter when both relatively few DMUs are available and environmental factors are used to restrict full comparison.

Another example is presented to illustrate a Culturally Adjusted DEA model to control the corporate culture's impacts on cross-bank branches' efficiency comparison. This model is applied to a real-life efficiency study of two major financial firms

Table 13.12 Results comparison (Paradi et al. 2010a)

	Separate frontier		Common frontier	
	Bank 1	Bank 2	Basic CCR	CA-DEA
Operational model (Bank 1 $SI = 0.43$, Bank 2 $SI = 0.55$)				
Efficiency (average)	0.88	0.85	0.80	0.84
SD	0.13	0.14	0.15	0.14
Median	0.92	0.86	0.83	0.86
Minimum	0.51	0.43	0.43	0.43
Profitability model (Bank 1 $CI = 0.93$, Bank 2 $CI = 0.54$)				
Efficiency (average)	0.90	0.93	0.87	0.90
SD	0.10	0.09	0.13	0.10
Median	0.92	0.97	0.90	0.92
Minimum	0.62	0.66	0.25	0.42

in Canada in 2000 (Paradi et al. 2010a). Two cultural indices were identified to represent two aspects of a firm's unique operating environment. The CI was designed to capture the nature and in turn the impact of a firm's corporate strategies, including resource allocation processes and product portfolios construction, and in turn to impact the branches' ability to coordinate their operating activities and optimize product diversity. The Service Capacity Index (SI) was calculated as a combined index of the branch's average daily hours of operation, branch size, branch age, and the number of Automated Teller Machines. It was assumed that these factors could affect a branch's overall service delivery capacity but were determined by the bank's top management.

The CI and SI can be estimated by using any number of suitable techniques, which depends on the specific research context. In the illustrative example in Paradi et al. (2010a), CI was estimated using the index number approach, which allowed for multilateral cross-sectional comparisons. SI was estimated by using a DEA model. CI was incorporated into the profitability model and SI was incorporated into the production model. The ratio form of the CA-CCR model with input orientation is formulated in (13.4), where I_j represents the cultural index score for unit j and I_0 is the cultural index of the unit under observation. Both I_j and I_0 are defined as ≤ 1.0.

$$\begin{aligned}
&\max_{u,v} \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \\
&\text{s.t.} \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq \max\left\{\frac{I_j}{I_o}, 1\right\}, \quad \text{for } j = 1, \ldots, n \\
&u_r, v_i \geq 0
\end{aligned} \tag{13.4}$$

Table 13.12 presents the results obtained from both the traditional and the Culturally Adjusted DEA model. The efficiencies obtained with respect to a single

bank's separate production frontier are also listed for reference purposes. The results confirm that the corporate cultural conditions play an important role in explaining the differences in branch efficiencies.

13.7 Validation

Managers are interested in improving their operations but whatever measurement methods are used to evaluate performance needs to be validated against other methods or even just against the managers' experience or expectations of how their world works. Formal validation is a critical success factor in having the results of an analysis accepted by those who are being measured. There are various approaches to validating results, and the following examples illustrate two on these approaches and may inspire the reader to devise their own strategies.

13.7.1 Validating a Method with Monte Carlo Simulation

A useful approach to validating DEA results, when appropriate, is to use statistical methods, such as Monte Carlo simulation. The CA-DEA model presented in Sect. 13.6.9 was used to adjust for cultural differences across two different banks (Paradi et al. 2010a). The goal was to eliminate corporation-specific influences on the branches being compared. The average efficiency differences between the branches from two banks become smaller after these adjustments. The question now is this: how valid is this approach?

To clear this concern, a Monte Carlo simulation analysis is conducted with the knowledge of the "true" production process and the existing inefficiencies. Two groups of DMUs (G1 and G2) are generated with the assumption that they have the same underlying production technology, but operate under different environments. A one output (Y)–three inputs (X_1, X_2, X_3) production function with constant returns to scale is specified in 13.5. The production is influenced by only one environmental factor (I), since this factor may be considered as an aggregated indicator representing all identifiable environmental factors. For each group, all inputs are generated from a random uniform distribution for 100 observations (representing DMUs) between the range (50, 150), and these values are fixed throughout the tests. There are 200 DMUs in total.

$$Y = (I^{\alpha} + \beta) \cdot \theta \cdot X_1^{0.5} X_2^{0.3} X_3^{0.2} \tag{13.5}$$

Following the procedure used in Ruggiero (1998), each DMU's true efficiency is calculated as $\theta = \exp(-|u|)$, where u is generated from a normal distribution, $N(0, 0.3)$. For each group, 15 observations are selected randomly and are deemed as efficient. The average efficiency is the same for both groups (0.83). The

environmental impacts on the estimated efficiency are tested by keeping the environmental indicator of the G1 group constant (1.0) while decreasing the indicator of the G2 group to values of 0.9, 0.67, and 0.5, corresponding to increased environmental differences of 10, 50, and 100% (relative to the indicator of the G2 group). The responsiveness of production to the environmental conditions is assumed to follow a non linear distribution, where α and β are parameters reflecting the responsiveness of production (assuming they are the same for all groups), $\alpha \geq 1$. In this example, $\alpha = 1.2$ and $\beta = 0$ for calculation convenience.

Since the objective here is to validate the model, whether the relationship between the environmental conditions and production and the values of some variables assigned here are realistic or not are not a concern.

The results confirm that the CA-DEA model can effectively compensate the branches operating under an unfavorable environment. Even when the environmental difference ratio increases to 2.0, the CA-DEA average efficiency is 0.80, which is not statistically significantly lower than the true average of 0.83 (at the 5% significance level), contrasting to 0.61 obtained from the traditional CCR model.

13.8 Conclusions

Banks are ubiquitous, operate both domestically in every country in the world as well as internationally, either directly (large banks from the developed countries) or indirectly through correspondent relationships. They have many branches, are typically well managed, and collect copious amounts of data on very detailed operational activities. Moreover, they are well accustomed to measuring performance and controlling operations by the very nature of what they do. For all of these reasons, banks offer an almost ideal study subject for DEA and many other analytic approaches.

Nevertheless, considerable caution is advisable when collecting data because, notwithstanding the large amounts available, there are many small and not so small errors in every dataset. So outliers may, in fact, be benefiting from data errors as easily as grossly inefficient DMUs may well be suffering from bad data. Also, different banks tend not to define data representing a certain activity or item in exactly the same manner, so "assets" may mean generally the same thing, but in fact, there are different components included depending on the banks' own decision or even by legislation, which complicates cross-bank or cross-country analyses.

Very different approaches must be taken when studying banks as the DMUs as opposed to the cases where the bank branches are the DMUs. This is naturally the case because the availability of data and the measures that matter to a bank, as an entity, or to a branch, as a unit, are very different. Much discussion also arises when defining the models with respect to which variables to include as inputs and which ones as outputs. Consideration must be given to both the technical issues involved and to the perception by management when results are delivered.

Many papers have been written on bank performance in many different domains of study, and more will be done in the future, as this is an interesting and challenging field to explore. DEA is a significant tool in this arena because it is nonparametric, does not need preconceived models, can be adapted to many views, is units independent (except for the Additive model), and allows for large models where the data is available. But it also has shortcomings, so the best approach is to use it in conjunction with other methodologies and validate the results against accepted processes in the financial services industry.

Acknowledgments We are indebted to Dr. Farvolden for proofreading the chapter and the former graduate students in the CMTE who had offered the results of some of their work.

References

Adler N, Golany B. Including principal component weights to improve discrimination in data envelopment analysis. J Oper Res Soc. 2002;53:985–91.

Aggarwall V. 1996, Performance Analysis of Large Canadian Banks over Time using DEA, MASc Dissertation, Centre for Management of Technology and Entrepreneurship.

Al-Faraj TN, Alidi AS, Bu-Bshait KA. Evaluation of bank branches by means of data envelopment analysis. Int J Oper Prod Man. 1993;13(9):45–53.

Al-Muharrami S. The causes of productivity change in GCC banking industry. Int J Prod Perform Manag. 2007;56:731–43.

Al-Sharkas AA, Hassan MK, Lawrence S. The impact of mergers and acquisitions on the efficiency of the US banking industry: further evidence. J Bus Finance Account. 2008;35:50–70.

Andersen P, Petersen NC. A procedure for ranking efficient units in data envelopment analysis. Manag Sci. 1993;39:1261–5.

Asmild M, Paradi JC, Aggarwall V, Schaffnit C. Combining DEA window analysis with the Malmquist index approach in a study of the Canadian banking industry. J Prod Anal. 2004;21:67–89.

Asmild M, Tam F. Estimating global frontier shifts and global Malmquist indices. J Prod Anal. 2007;27:137–48.

Athanassopoulos AD. Service quality and operating efficiency synergies for management in the provision of financial services: evidence from Greek bank branches. Eur J Oper Res. 1997;98:300–13.

Athanassopoulos AD. Non-parametric frontier models for assessing the market and cost efficiency of large-scale bank branch networks. J Money Credit Bank. 1998;30(2):172–92.

Athanassopoulos AD. An optimization framework of the triad: service capabilities, customer satisfaction and performance. In: Harker PT, Zenios SA, editors. Performance of financial institutions. Cambridge, England: Cambridge University Press; 2000. p. 312–35.

Avkiran NK. The evidence on efficiency gains: the role of mergers and the benefits to the public. J Bank Finance. 1999;23:991–1013.

Avkiran NK. Productivity Analysis in the Services Sector with Data Envelopment Analysis, 3rd ed., University of Queensland Business School, The University of Queensland, Brisbane 2006.

Avkiran NK. Removing the impact of environment with units-invariant efficient frontier analysis: an illustrative case study with intertemporal panel data. Omega. 2009;37:535–44.

Banker RD, Morey R. The use of categorical variables in data envelopment analysis. Manage Sci. 1986;32(12):1613–27.

Bauer PW. Recent developments in the econometric estimation of frontiers. J Econometrics. 1990;46:39–56.
Bauer PW, Berger AN, Ferrier GD, Humphrey DB. Consistency conditions for regulatory analysis of financial institutions: a comparison of frontier efficiency methods. J Econ Bus. 1998;50:85–114.
Berg SA, Førsund FR, Jansen ES. Malmquist indices of productivity growth during the deregulation of Norwegian banking, 1980–1989. Scand J Econ (Suppl). 1992;94:211–28.
Bergendahl G, Lindblom T. Evaluating the performance of Swedish savings banks according to service efficiency. Eur J Oper Res. 2008;185:1663–73.
Berger AN, DeYoung R. Problem loans and cost efficiency in commercial banks. J Bank Finance. 1997;21:849–70.
Berger AN, Humphrey DB. 1990, Measurement and Efficiency Issues in Commercial Banking, Finance and Economic Discussion Series, Working Paper #151, Board of Governors of the Federal Reserve System.
Berger AN, Humphrey DB. Efficiency of financial institutions: international survey and directions for future research. Eur J Oper Res. 1997;98:175–212.
Berger AN. The efficiency effects of bank mergers and acquisitions: a preliminary look at the 1990s data. In: Amihud Y, Miller G, editors. Bank mergers and acquisitions. Dordrecht: Kluwer; 1998. p. 79–111.
Bhattacharyya A, Lovell CAK, Sahay P. The impact of liberalization on the productive efficiency of Indian commercial banks. Eur J Oper Res. 1997;98:333–46.
Camanho AS, Dyson RG. Efficiency, size, benchmarks and targets for bank branches: an application of data envelopment analysis. J Oper Res Soc. 1999;50:903–15.
Canbas S, Cabuk A, Kilic SB. Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case. Eur J Oper Res. 2005;166:528–46.
Casu B, Molyneux P. A comparative study of efficiency in European banking. Appl Econ. 2003;35:1865–76.
Caves DW, Christensen LR, Diewert WE. The economic theory of index numbers and the measurement of input, output and productivity. Econometrica. 1982;50:1393–414.
Cazals C, Florens JP, Simar L. Nonparametric frontier estimation: a robust approach. J Econometrics. 2002;106:1–25.
Coelli T, Rao DSP, Battese GE. An introduction to efficiency and productivity analysis. Boston, MA: Kluwer; 1998.
Colwell RJ, Davis EP. 1992, Output and Productivity in Banking, Scandinavian Journal of Economics.
Cook WD, Hababou M, Tuenter HJ. Multicomponent efficiency measurement and shared inputs in data envelopment analysis: an application to sales and service performance in bank branches. J Prod Anal. 2000;14:209–24.
Cook WD, Seiford LM, Zhu J. Models for performance benchmarking: measuring the effect of e-business activities on banking performance. Omega. 2004;32:313–22.
Cooper WW, Seiford LM, Tone K. Data envelopment analysis: a comprehensive text with models, applications, references, and DEA-solver software. New York: Springer; 2007.
Das A, Ghosh S. Financial deregulation and efficiency: an empirical analysis of Indian banks during the post reform period. Rev Financ Econ. 2006;15:193–221.
Delis MD, Papanikolaou NI. Determinants of bank efficiency: evidence from a semi-parametric methodology. Manag Finance. 2009;35:260–75.
Deville A. Branch banking network assessment using DEA: a benchmarking analysis – a note. Manag Acc Res. 2009;20:252–61.
DeYoung R. 1993, Determinants of Cost Efficiencies in Bank Mergers, Economic and Policy Analysis, Working Paper 93–01, Washington: Office of the Comptroller of the Currency.
Drake L, Hall MJB. Efficiency in Japanese banking: an empirical analysis. J Bank Finance. 2003;27:891–917.

Drake L, Howcroft B. Relative efficiency in the branch network of a UK bank: an empirical study. Omega. 1994;22(1):83–91.
Drake L, Hall MJB, Simper R. The impact of macroeconomic and regulatory factors on bank efficiency: a non-parametric analysis of Hong Kong's banking system. J Bank Finance. 2006;30:1443–66.
Favero C, Papi L. Technical efficiency and scale efficiency in the Italian banking sector: a non-parametric approach. Appl Econ. 1995;27:385–95.
Färe R, Grosskopf S, Norris M, Zhang Z. Productivity growth, technical progress, and efficiency changes in industrialized countries. Am Econ Rev. 1994;84:66–83.
Färe R, Grosskopf S. Nonparametric productivity analysis with undesirable outputs: comment. Am J Agric Econ. 2003;85:1070–4.
Ferrier GD, Lovell CAK. Measuring cost efficiency in banking: econometric and linear programming evidence. J Econometrics. 1990;46:229–45.
Fethi MD, Pasiouras F. Assessing bank efficiency and performance with operational research and artificial intelligence techniques: a survey. Eur J Oper Res. 2010;204:189–98.
Fixler D, Zieschang K. 1999, The Productivity of the Banking Sector: Integrating Financial and Production Approaches to Measuring Financial Service Output, Canadian Journal of Economics, 32, No. 2.
Florens JP, Simar L. Parametric approximations of nonparametric frontiers. J Econometrics. 2005;124:91–116.
Fox KJ, Hill RJ, Diewert WE. Identifying outliers in multi-output models. J Prod Anal. 2004;22:73–94.
Frei FX, Harker PT. Measuring aggregate process performance using AHP. Eur J Oper Res. 1999;116:436–42.
Fukuyama H, Weber WL. Japanese banking inefficiency and shadow pricing. Math Comput Model. 2008;48:1854–67.
Giokas D. Bank branch operating efficiency: a comparative application of DEA and the log-linear model. Omega. 1991;19(6):549–57.
Golany B, Storbeck JE. A data envelopment analysis of the operation efficiency of bank branches. Interfaces. 1999;29(3):14–26.
Grabowski R, Mathur I, Rangan N. The role of takeovers in increasing efficiency. Manag Decis Econ. 1995;16(3):211–24.
Hauner D. Explaining efficiency differences among large German and Austrian banks. Appl Econ. 2005;37:969–80.
Havrylchyk O. Consolidation of the Polish banking sector: consequences for the banking institutions and the public. Econ Syst. 2004;28:125–40.
Haynes M, Thompson S. The productivity effects of bank mergers: evidence from the UK building societies. J Bank Finance. 1999;23:825–46.
Ho CT, Wu YS. Benchmarking performance indicators for banks. Benchmarking Int J. 2006;13:147–59.
Jablonsky J, Fiala P, Smirlis Y, Despotis DK. DEA with interval data: an illustration using the evaluation of branches of a Czech bank. CEJOR. 2004;12:323–37.
Jenkins L, Anderson M. Multivariate statistical approach to reducing the number of variables in data envelopment analysis. Eur J Oper Res. 2003;147:51–61.
Kantor J, Maital S. Measuring efficiency by product group: integrating DEA with activity-based accounting in a large Mideast bank. Interfaces. 1999;29(3):27–36.
Kim CS, Davidson LF. The effects of IT expenditures on banks' business performance: using a balanced scorecard approach. Manag Finance. 2004;30:28–45.
Kinsella RP. The measurement of bank output. J Inst Bankers Ireland. 1980;82:173–83.
Liu J, Tone K. A multistage method to measure efficiency and its application to Japanese banking industry. Soc Econ Plann Sci. 2008;42:75–91.
Liu ST. Measuring and categorizing technical efficiency and productivity change of commercial banks in Taiwan. Expert Syst Appl. 2010;37:2783–9.

Lovell CAK, Pastor JT. Target setting: an application to a bank branch network. Eur J Oper Res. 1997;98:290–9.
Lozano-Vivas A, Pastor JT, Pastor JM. An efficiency comparison of European banking systems operating under different environmental conditions. J Prod Anal. 2002;18:59–77.
Malmquist S. Index numbers and indifference surfaces. Trabajos de Estatistica. 1953;4:209–42.
McEachern D, Paradi JC. Intra- and inter-country bank branch assessment using DEA. J Prod Anal. 2007;27:123–36.
Mukherjee K, Ray SC, Miller SM. Productivity growth in large US commercial banks: the initial post-deregulation experience. J Bank Finance. 2001;25:913–39.
Nash D, Sterna-Karwat A. An application of DEA to measure branch cross selling efficiency. Comput Oper Res. 1996;23(4):385–92.
Oliveira C, Tabak B. An international comparison of banking sectors: a DEA approach. Global Econ Rev. 2005;34:291–307.
Oral M, Kettani O, Yolalan R. 1992, An Empirical Study on Analyzing the Productivity of Bank Branches, IIE Transactions 24 No 5.
Oral M, Yolalan R. An empirical study on measuring operating efficiency and profitability of bank branches. Eur J Oper Res. 1990;46:282–94.
Paradi JC, Schaffnit C. Commercial branch performance evaluation and results communication in a Canadian bank – a DEA application. Eur J Oper Res. 2004;156:719–35.
Paradi JC, Zhu H, Edelstein B. 2009, Identifying management groups in a large Canadian bank branch network with DEA approach, XI European Workshop on Efficiency and Productivity Analysis, Pisa, Italy.
Paradi J, Vela S, Zhu H. Adjusting for cultural differences, a new DEA model applied to a merged bank. J Prod Anal. 2010a;33:109–23.
Paradi JC, Rouatt S, Zhu H. 2010b, Two-Stage Evaluation of Bank Branch Efficiency Using Data Envelopment Analysis. Accepted for publication in Omega International Journal of Management Science.
Parkan C. Measuring the efficiency of service operations: an application to bank branches. Eng Costs Prod Econ. 1987;12:237–42.
Pasiouras F. International evidence on the impact of regulations and supervision on banks' technical efficiency: an application of two-stage data envelopment analysis. Rev Quant Finance Acc. 2008;30:187–223.
Pastor J, Perez F, Quesada J. Efficiency analysis in banking firms: an international comparison. Eur J Oper Res. 1997;98:396–408.
Peristiani S. 1993, Evaluating The Postmerger X-Efficiency and Scale Efficiency of US Banks, Federal Reserve Bank Of New York.
Peristiani S. Do mergers improve the X-efficiency and scale efficiency of US banks? Evidence from the 1980s. J Money Credit Bank. 1997;29(3):326–37.
Portela MCAS, Thanassoulis E. Comparative efficiency analysis of Portuguese bank branches. Eur J Oper Res. 2007;177:1275–88.
Ray SD. Data envelopment analysis: theory and techniques for economics and operations research. Cambridge, UK: Cambridge University Press; 2004.
Resti A. Evaluating the cost-efficiency of the Italian banking system: what can be learned from the joint application of parametric and non-parametric techniques. J Bank Finance. 1997;21:221–50.
Resti A. Regulation can foster mergers, can mergers foster efficiency? The Italian case. J Econ Bus. 1998;50:157–69.
Ruggiero J. Non-discretionary inputs in data envelopment analysis. Eur J Oper Res. 1998;111:461–9.
Sathye M. Measuring productivity changes in Australian banking: an application of Malmquist indices. Manag Finance. 2002;28:48–59.
Schaffnit C, Rosen D, Paradi JC. Best practice analysis of bank branches: an application of DEA in a large Canadian bank. Eur J Oper Res. 1997;98(2):269–89.

Seçme NY, Bayrakdaroğlu A, Kahraman C. Fuzzy performance evaluation in Turkish banking sector using analytic hierarchy process and TOPSIS. Expert Syst Appl. 2009;36:11699–709.
Seiford LM, Zhu J. Profitability and marketability of the top 55 US commercial banks. Manage Sci. 1999;45:1270–88.
Sherman HD, Ladino G. Managing bank productivity using data envelopment analysis (DEA). Interfaces. 1995;25(2):60–73.
Sherman HD, Gold F. Bank branch operating efficiency: evaluation with data envelopment analysis. J Bank Finance. 1985;9(2):297–316.
Sherman HD, Rupert TJ. Do bank mergers have hidden or foregone value? Realized and unrealized operating synergies in one bank merger. Euro J Oper Res. 2006;168:253–68.
Sherman HD, Zhu J. Benchmarking with quality-adjusted DEA (Q-DEA) to seek lower-cost high-quality service: Evidence from a US bank application. Ann Oper Res. 2006;145:301–19.
Simar L. Detecting outliers in frontier model: a simple approach. J Prod Anal. 2003;20:391–424.
Smith P. Data envelopment analysis applied to financial statements. Omega Int J Manage Sci. 1990;18(2):131–8.
Soteriou AC, Stavrinides Y. An internal customer service quality data envelopment analysis model for bank branches. Int J Oper Prod Manage. 1997;17(8):780–9.
Soteriou A, Zenios SA. Operations, quality, and profitability in the provision of banking services. Manage Sci. 1999;45(9):1221–38.
Soteriou AC, Stavrinides Y. An internal customer service quality data envelopment analysis model for bank branches. Int J Bank Mark. 2000;18:246–52.
Sowlati T, Paradi JC. Establishing the "practical frontier" in data envelopment analysis. Omega. 2004;32:261–72.
Stanton KR. Trends in relationship lending and factors affecting relationship lending efficiency. J Bank Finance. 2002;26:127–52.
Stavárek D. Banking efficiency in the context of European integration. East Eur Econ. 2006;44:5–31.
Sufian F, Majid MZA. Bank mergers performance and the determinations of Singaporean banks' efficiency: an application of two-stage banking models. Gadjah Mada Int J Bus. 2007;9:19–39.
Thompson RG, Brinkmann EJ, Dharmapala PS, Gonzales-Lima MD, Thrall RM. DEA/AR profit ratios and sensitivity of 100 large US banks. Eur J Oper Res. 1997;98:213–29.
Thoraneenitiyan N, Avkiran NK. Measuring the impact of restructuring and country-specific factors on the efficiency of post-crisis East Asian banking systems: integrating DEA with SFA. Socio Econ Plann Sci. 2009;43:240–52.
Tortosa-Ausina E. Exploring efficiency differences over time in the Spanish banking industry. Eur J Oper Res. 2002;139:643–64.
Tortosa-Ausina E, Grifell-Tatje E, Armero C, Conesa D. Sensitivity analysis of efficiency and Malmquist productivity indices: an application to Spanish savings banks. Eur J Oper Res. 2008;184:1062–84.
Tulkens H. On FDH efficiency analysis: some methodological issues and applications to retail banking, courts and urban transit. J Prod Anal. 1993;4(1/2):183–210.
Vance H. 2000, Opportunity Index Development for Bank Branch Networks, MASc Dissertation, Centre for Management of Technology and Entrepreneurship.
Vander Vennet R. The effect of mergers and acquisitions on the efficiency and profitability of EC credit institutions. J Bank Finance. 1996;20:1531–58.
Vassiloglou M, Giokas D. A study of the relative efficiency of bank branches: an application of data envelopment analysis. J Oper Res Soc. 1990;41(7):591–7.
Vela S. 2002, Measuring Effects Of Cultural Differences In Bank Branch Performance And Implications To The Branch Integration Process In Bank Mergers, Ph.D. Thesis, Centre for Management of Technology and Entrepreneurship.
Webb RM. Levels of efficiency in UK retail banks: A DEA window analysis. Int J Econ Bus. 2003;10:305–22.

Weill L. Measuring cost efficiency in European banking: a comparison of frontier techniques. J Prod Anal. 2004;21:133–52.
Wu S. Bank mergers and acquisitions – an evaluation of the 'four pillars' policy in Australia. Aust Econ Pap. 2008;47:141–55.
Wu HY, Tzeng GH, Chen YH. A fuzzy MCDM approach for evaluating banking performance based on balanced scorecard. Expert Syst Appl. 2009;36:10135–47.
Yang Z, Paradi JC. 2006, Cross Firm Bank Branch Benchmarking Using "Handicapped" Data Envelopment Analysis to Adjust for Corporate Strategic Effects, System Sciences, Proceedings of the 39th Annual Hawaii International Conference 2, 34b–34b.
Yavas BF, Fisher DM. Performance evaluation of commercial bank branches using data envelopment analysis. J Bus Manage. 2005;11:89–102.
Zenios CV, Zenios SA, Agathocleous K, Soteriou AC. Benchmarks of the efficiency of bank branches. Interfaces. 1999;29(3):37–51.
Zopounidis C, Despotis DK, Stavropoulou E. Multiattribute evaluation of Greek banking performance. Appl Stoch Models Data Anal. 1995;11:97–107.

Chapter 14
Engineering Applications of Data Envelopment Analysis

Issues and Opportunities

Konstantinos P. Triantis

Abstract Engineering is concerned with the design of products, services, processes, or in general with the design of systems. These design activities are managed and improved by the organization's decision-makers. Therefore, the performance evaluation of the production function where engineering plays a fundamental role is an integral part of managerial decision-making. In the last 20 years, there has been limited research that uses data envelopment analysis (DEA) in engineering. One can attribute this to a number of issues that include but are not limited to the lack of understanding of the role of DEA in assessing and improving design decisions, the inability to open the input/output process transformation box, and the unavailability of production and engineering data. Nevertheless, the existing DEA applications in engineering have focused on the evaluation of alternative design configurations, have proposed performance improvement interventions for production processes at the disaggregated level, assessed the performance of hierarchical manufacturing organizations, studied the dynamical behavior of production systems, and have dealt with data imprecision issues. This chapter discusses the issues that the researcher faces when applying DEA to engineering problems, proposes an approach for the design of an integrated DEA-based performance measurement system, summarizes studies that have focused on engineering applications of DEA, and suggests some systems thinking concepts that are appropriate for future DEA research in engineering.

Keywords Data envelopment analysis • Engineering design and decision-making • Disaggregated process definition and improvement • Hierarchical manufacturing system performance • Dynamical production systems • Data imprecision • Integrated performance measurement systems • Systems thinking

K.P. Triantis (✉)
Virginia Tech, System Performance Laboratory, Grado Department of Industrial and Systems Engineering, Northern Virginia Center, 7053 Haycock Road, Falls Church, VA 22043-2311, USA
e-mail: triantis@vt.edu

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_14,

14.1 Background and Context

Investments are made by the engineering organization to enhance its ability to provide quality products/services to its customers in the most efficient and effective manner. Investments are typically made in facilities and equipment and for process improvement programs. Consequently, the engineering organization's investment and production functions are not only interlinked but also inseparable. A part of the domain of the investment function is the provision of capital for projects at the benchmark or better rate of return. On the other hand, the production function is primarily involved with the design of the production processes or systems that provide the organization's products and services. As part of the production function, design engineers define and introduce product and service concepts based on customer requirements and process engineers design and improve production processes.

Consequently, the organization needs to be successful on both fronts, i.e., in managing the investment and production function activities concurrently. Well-known criteria such as net present worth, internal rate of return, benefit/cost ratios, etc. are used to assess the organization's success in carrying out its investment activities. Additionally, operational performance is typically assessed using a different set of criteria such as efficiency, productivity, quality, outcome performance, etc. Nevertheless, investments in equipment and products typically have not been evaluated using efficiency-based operational criteria. For notable exceptions see Bulla et al. (2000) and Sun (2002).

One can make the assertion that engineering is primarily concerned with the design of products, services, processes, or in general with the design of systems. Furthermore, these design activities are managed and improved by the organization's decision-makers. Therefore, the performance evaluation of the production function where engineering plays a fundamental role is an integral part of managerial decision-making.

Nevertheless, even though engineers have always been involved with the design and improvement of the production function, in academia, the study of the production function performance has been primarily the domain of economists and operations research analysts. Surprisingly, very few engineering curricula havecourses and/or research programs dedicated to the teaching and study of productive performance.

On the other hand, it should be noted that the measurement and evaluation of the efficiency performance of production processes is what initially provided the impetus for the theoretical developments of Debreu (1951), Koopmans (1951), and Shephard (1953, 1970). However, many of the first and subsequent analyses of efficiency performance have focused on entire economies or sectors of the economy.

Therefore, the empirical evaluation of the disaggregated efficiency performance of production processes has not been pursued as rigorously as the efficiency

evaluations that focus on the economy, industries, or firms. This is not necessarily surprising given that traditional cost and financial accounting systems rarely define and collect data that would be useful for process-related studies. On the other hand, engineers typically track alternative performance measures in inventory management, product quality, and customer order management. Nevertheless, the relationship of these engineering-based measures to the Farrell (1957), Koopmans (1951), and Charnes et al. (1978) representations of efficiency performance have neither been extensively analyzed nor been documented.

Also, within engineering, the linkage of performance measurement to process performance improvement is of great concern. In practice, performance improvement requires that the engineer identify the causal factors that are associated with efficiency performance. However, Färe et al. (1994) argue that the measurement of efficiency performance is by itself an appropriate and necessary research endeavor. They continue by stating that the investigation of the causes of efficiency or inefficiency is important but that the measurement of efficiency by itself can provide some important insights, i.e., discovering the patterns of efficiency performance without hypothesizing about the causal factors.

Causality and process improvement have been critical points of departure from what engineers are practically involved with on a daily basis and at least initially what researchers in efficiency and productivity performance literature have not shown a great deal of interest in. This perhaps is one of the reasons why there has been such a gap between the research in efficiency and productivity and the practice of engineering. However, it is encouraging to observe that recently the efficiency literature is actively pursuing the issues of causality and process improvement.

Consequently, even though efficiency measurement has been a fruitful area of scholarly research, decision-makers and engineers have not extensively implemented efficiency measurement concepts to improve the design performance of products and the transformation performance of production processes. The challenge remains to focus on the measurement of product design alternatives and disaggregated process performance and evaluate the appropriateness of the nonparametric and parametric approaches (Fried et al. (1993)) that have been proposed in the literature insofar as to their leading to improved decision making resulting in product and process improvements.

Given that data envelopment analysis (DEA) (Charnes et al. (1978)) research does not have a well-defined role in the various engineering disciplines, the National Science Foundation provided funding for a workshop that was held at Union College in December of 1999 to explore issues and applications of DEA in engineering. A number of engineers from Union College attended the workshop where they were exposed to efficiency measurement and DEA (Seiford 1999) and to some existing DEA engineering applications. These applications involved road maintenance (Cook et al. 1990, 1994), turbofan jet engines (Bulla et al. 2000), electric power systems (Criswell and Thompson 1996), and circuit board manufacturing (Otis 1999; Hoopes et al. 2000; Hoopes and Triantis 2001; Triantis and Otis 2003). During the workshop small breakout sessions explored

research issues and attempted to define interesting engineering applications of DEA forfurther study. One of the key issues that arose during the workshop was how well-established engineering design methodologies such as design of experiments (DOE) could effectively interface with the DEA methodology. One of the fundamental conclusions of the workshop is the need to explore these methodological interfaces and to continue to pursue innovative engineering DEA applications.

Additionally, as part of recent productivity and efficiency analysis workshops starting with the tenth EWEPA conference in 2007, continuing with the 2008 NAPW, and the 11th EWEPA conference in 2009, roundtable sessions were dedicated to the topic of engineering applications of efficiency analysis. These sessions were facilitated by A. Johnson of Texas A&M and Kostas Triantis from Virginia Tech. The issues covered included:

1. The evaluation of production axioms and hypotheses in the context of the design of engineering systems.
2. Lack of guidance in the definition of what constitutes a reasonable unit of analysis especially when systems are unique and have not been designed before.
3. Lack of guidance in specifying reasonable input, output, and environmental representations over the life-cycle of the system.
4. The lack of an apparent connection between the underlying technologies, the transformation process(es), and the defined variables.
5. Information systems that are not typically designed to facilitate operational performance analyses lead to difficulties in gathering and using data.
6. The linkage to decision making is ambiguous and the mapping between the modeling world and the "real world" is not obvious.

The specific applications that were discussed included the following: road maintenance operations, transportation network performance, warehousing operations, portfolio algorithms, satellite scheduling among others.

The objectives of this chapter are to present some of the fundamental research issues and opportunities when exploring engineering applications of DEA (Sect. 14.2), to propose an approach that can be used to design an integrated DEA-based performance measurement system (Sect. 14.3), to summarize selected engineering DEA applications (Sect. 14.4), to discuss some systems thinking concepts that can guide future DEA research in engineering (Sect. 14.5), and to provide an extensive bibliography (Sect. 6) of the engineering applications of DEA. This reference list includes refereed papers, Ph.D. dissertations, and M.S. theses. It should be noted that even though the coverage in this chapter is intended to be as comprehensive as possible, there are references that have been inadvertently omitted. The author apologizes to the authors of the references that have not been included and would appreciate if they sent him an e-mail stating that a reference has been omitted so as to include these references in a possible subsequent edition of this handbook.

14.2 Research Issues and Opportunities

The intent of this section is to concisely discuss some of the issues and opportunities faced by researchers when studying the engineering applications of DEA. These issues are by no means the only ones that researchers face when engaged in this research domain. However, they can be viewed as a means to establish a background on which future discussions on these and other issues can occur. Finally, the sequence of the issues presented in this section is by no means an indication of their relative importance.

14.2.1 Evaluating Design Alternatives

As stated earlier, the primary concern of engineering is design. The fundamental question is how can DEA provide assistance for engineers that choose among different design alternatives? Within the context of this question lie three other fundamental questions.

Can DEA be used effectively with other methodologies such as DOE to assist in establishing better product design configurations? On the surface, there is no reason why this research direction should not to be pursued. Cooper (1999) states that DEA has been used in conjunction with other disciplines and/or approaches. The fundamental provision is that DEA provides important additional insights for the analyst, engineer, and decision-maker. This remains a conjecture that needs to be ascertained for each DEA research study that combines different methodologies.

Can DEA be used to contribute to the design of effective process improvement interventions? The answer to this question is positive as long as the specification of the input/output variables accurately represents the underlying production processes. Farrell (1957) discussed that although there are many possibilities in evaluating productive performance, two at once suggest themselves – a theoretical function specified by engineers[1] and an empirical function based on the best results observed in practice.

The former would be a very natural concept to choose since a theoretical function represents the best that is theoretically obtainable. Certainly, it is a concept used by engineers themselves when they discuss the efficiency of a machine or process. However, although the theoretical function is a reasonable and perhaps the best concept for the efficiency of a single production process, there are considerable objections to its application to anything so complex as a

[1] There have been attempts in the literature to define the theoretical production function. However, the focus of this chapter is to build on the notion of the empirical production function and how it has been used in engineering applications.

typical manufacturing firm, let alone an industry. Nevertheless, one can argue that in order to make any reasonable suggestion for a process improvement intervention there is a need for the input/output variables of the empirical function to effectively represent the underlying production processes, i.e., accurately denote how inputs are transformed into outputs. Cook et al. (1990, 1994) and Hoopes and Triantis (2001) document examples where this is accomplished.

Is the DEA model valid? If it is not, then the researcher will be hard pressed to recommend the DEA results as a means to choose among alternative product and process design configurations. There have been attempts in the literature to investigate the issue of pitfalls/protocols (Dyson et al. (2001)) and quality (Pedraja-Chaparro et al. 1999) of the DEA analysis. However, researchers need to investigate further what is meant by model validity in DEA. This should be pursued in conjunction with an understanding of the relationship between the virtual representation of the DEA model and the "real" world (see Sect. 5.3 of this handbook).

14.2.2 Disaggregated Process Evaluation and Improvement: Opening the "Input/Output Transformation Box"

As stated earlier, even though efficiency measurement has been a fruitful area of scholarly research, decision-makers have not used the DEA framework (Charnes et al. (1978)) to evaluate and improve the performance of production/manufacturing processes. A notable exception is provided by the work of Cooper et al. (1996).

Consequently, the challenge remains on how to focus on disaggregated production processes, i.e., how to study in detail the "input/output transformation box." This entails understanding the technologies used, the processes employed, the organizational structures, and the decision-making rules and procedures. An important modeling and research issue for one who decides to pursue this line of research is to decide the appropriate level of aggregation (or the necessary level of detail complexity) of the DEA analysis.

One of the main reasons for the lack of efficiency studies at the disaggregated production process level is that one needs to invest a considerable amount of time within organizations studying their transformation processes, interviewing both decision-makers and employees, and collecting useful data and information. This means that the researcher is interested in acquiring a deep appreciation of thephysical, organizational, and decision-making structures within the organization. By understanding these structures the researcher will also understand the performance behavior of the organization and consequently gain an in-depth appreciation of all performance measurement issues. To gain access to organizations, decision-makers need to accept that the DEA approach can potentially lead to significant organizational learning and important improvements in performance.

Once within the organization, the researcher may not have the data to support a complete model where all the input and output variables are included. Nevertheless, formulations that attempt to capture performance purely from the input or the output perspective can lead to potentially useful insights. This is important when there is an effort to incorporate alternative concepts of "inputs" and "outputs" in performance evaluation such as process and product characteristics. This notion of focusing on performance evaluation models solely from the input or the output perspective is consistent with the unified framework for classifying DEA models presented by Adolphson et al. (1990). Nevertheless, there is a need for academic research to emphasize innovative input/output specifications that represent the underlying processes and structures whose performance is being evaluated.

One of the DEA modeling approaches that has been proposed to measure and improve disaggregated process performance is the network model (Färe and Grosskopf (2000)). The fundamental concept behind this approach is for the researcher to open the "input/output transformation box" represented by the DMU meaning, as stated earlier, that the researcher is concerned with how the transformation of inputs into outputs is accomplished. For example, intermediate outputs of one stage may be used as inputs into another stage. One can proceed to evaluate the efficiency performance of the specific production stages (represented by nodes) as well as the overall performance of the DMU. In addition to evaluating the relationship of different production stages, the analyst can also evaluate the performance of a production node and the performance of a customer node (Löthgren and Tambour (1999)) that are interrelated. This approach facilitates the allocation of the resources between the customer-oriented activities and the traditional production activities. In this way, one can also study the satisfaction of different goals associated with each node.

14.2.3 Hierarchical Manufacturing System Performance[2]

In addition to the study of disaggregated efficiency performance, DEA research can focus performance management decision making that coordinates activities among different levels in a manufacturing organization. Nijkamp and Rietveld (1981) present three principal problems of policy making in multilevel environments: Interdependencies among the components of the system; conflicts among various priorities, objectives, and targets *within* individual components of the system; conflicts among priorities, objectives, and targets *among* the various components of the system.

[2] Part of the material of this section is adopted from Hoopes, B., Triantis, K., and N. Partangel, 2000, The Relationship Between Process and Manufacturing Plant Performance: A Goal Programming Approach, *International Journal of Operations and Quantitative Management*, 6(4), 287–310.

In the context of manufacturing environments analogies can be found for each of these three principal problems. Clearly, technological and administrative interdependencies exist among the various production processes. Within a single production process, conflicts exist with respect to objectives; for example, quality assurance and throughput objectives are usually at odds, especially in the short term. Finally, conflicts with respect to manufacturing performance targets in terms of throughput, cost, efficiency, allocation of resources, etc. exist among the various manufacturing departments.

Furthermore, the achievements of the production-planning goals at different time periods in the planning cycle are complementary. The achievement of the overall plant-level manufacturing targets usually depends on the performance of individual processes that provide the intermediate products. Consequently, one would expect that the improvement of the operational performance of the individual manufacturing processes at specific time periods should favorably impact the overall plant performance of the manufacturing facility for the entire planning cycle. This constitutes a fundamental research hypothesis that needs to be explored further.

Based on the three previously mentioned problems associated with multilevel organizations, Nijkamp and Rietveld (1981) suggest the usefulness of multiobjective analytical techniques such as goal programming for addressing planning problems. However, within the efficiency measurement field, the use of goal programming in conjunction with DEA has not been studied extensively. Efficiency performance has been linked with other performance dimensions such as quality and effectiveness but only rarely with respect to accomplishment of conflicting and/or complementary performance goals. Goal programming is an analytical approach that allows for the explicit incorporation of these types of goals. DEA provides performance assessments with respect to resource utilization and/or output/outcome achievement. Therefore, the combination of goal programming with DEA allows the analysis to focus on goal achievement within an organizational hierarchy and at the same time address performance measurement with respect to resource utilization and output production. Given the nature of the production planning function in most manufacturing organizations, a goal programming DEA approach is appropriate when assessing organizational performance.

Thanassoulis and Dyson (1992) and Athanassopoulos (1995) have introduced goal programming as means of assessing whether organizational input/output targets are met. Thanassoulis and Dyson's (1992) formulation provides a method to estimate the input/output targets for each individual decision-making unit (DMU) in an organization. However, it does not address the planning and resource allocation issues at the global organizational level while considering all the DMUs simultaneously. This means that their formulation does not include global organizational targets or global organizational constraints. Athanassopoulos (1995) provided these enhancements with his formulation of a Goal Programming and Data Envelopment Analysis (GODEA) approach that combines conflicting objectives of efficiency, effectiveness, and equity in resource allocation for service organizations.

It should be noted that the extension of the GODEA approach to manufacturing environments requires a fundamental rethinking of how to incorporate the decisions made at the different levels within the manufacturing hierarchy. Process supervisors are primarily concerned with attaining the required throughput that is defined by the master production plan. Process engineers are concerned with well-balanced production processes that minimize bottlenecks and reduce idle time. Plant managers require that plant level output requirements be met so as to ensure that demand requirements are satisfied. Above all, management is interested in the overall performance of the manufacturing facility and the extent to which all of these previously defined goals are aligned.

Additionally, within a manufacturing context, the goal programming analytical approach needs to define what constitutes an appropriate unit of analysis, labeled as a DMU in DEA. At the very minimum the goal programming approach should provide information as to how realistic the process and plant level targets are, should identify processes that are under-performing, and should point to serious bottlenecks. This is considerably more informative than one would obtain through the completion of a typical DEA evaluation of each production process and of the plant as a whole.

Note that a fundamental assumption of these modeling approaches is that each of the processes in a manufacturing facility is uniquely different from a technological point of view. This implies that the operational performance of a single process cannot be used to benchmark other processes in the manufacturing facility. Consequently, each process is compared to itself over the planning horizon, and one must evaluate the performance of each production process independently, while recognizing the process interdependencies by linking adjacent processes through their respective output levels.

Based on the need to align overall plant and production process goals, Hoopes et al. (2000) assess the contribution of the operational performance of individual production processes at specific time periods for the achievement of the overall plant production-planning targets for the entire production planning cycle. They present a goal programming DEA formulation that evaluates the achievement of plant and production process targets for a manufacturing technology that involves serial production stages. They then illustrate their approach by evaluating the process and plant effectiveness and serial production performance of a circuit board manufacturing facility for which plant and process level data have been accumulated over a 2-year time horizon. The results obtained from the proposed goal programming approach are compared to the radial variable returns to scale efficiency scores (Banker et al. (1984)) in terms of evaluating the overall performance of the manufacturing facility.

In this research, the concepts of effectiveness and balance are addressed in a manufacturing context. The production-planning function predetermines the goals or targets both at the plant and production process levels. Process effectiveness is the degree to which a production process meets its resource (input) and throughput (output) targets for a specific time period. Plant-level effectiveness is the degree to which the manufacturing facility meets its plant-level production-planning targets

over the entire planning period. Finally, the concept of production line balance takes into account the serial nature of the production processes so as to evaluate the technology's production bottlenecks. Note that by substituting actual input and output levels, the model proposed in this research can be used to also evaluate the performance of the production processes in terms of their resource utilization and output production process efficiency.

14.2.4 Data Measurement Imprecision in Production Systems[3]

Whether one uses a frontier (DEA, stochastic frontier (Aigner et al. 1977; Battese and Corra 1977; Meeusen and Vanden Broeck 1977) or a nonfrontier approach (Free Disposal Hull (Deprins et al. 1984), pair-wise dominance (Koopmans 1951), data measurement imprecision is more the norm rather than the exception for production and manufacturing systems.

Production data are usually not gathered for the sole purpose of conducting production analyses. The problem is further compounded by the fact that there are multiple measurement systems in place that account for different segments of the production process (Otis (1999)). These measurement systems are rarely consistent in the way they collect information and in their definitions of the accounts associated with production processes.

Furthermore, in most of the real-world decision-making problems, the data are not always known *precisely* or the information regarding certain parameters that are part of a mathematical model is not readily available. For example, when evaluating the performance of urban transit alternatives, imprecision exists whenmeasuring the headway between vehicles.

The "classical" DEA formulation lacks the flexibility to deal with *imprecise* data or data expressed in *linguistic form*. This problem can be approached by expanding the current DEA technology to allow imprecision in the ordinal rankings and the knowledge only of bounds of certain parameters (Cooper et al. 2000) or with the use of fuzzy set theory (Zadeh (1965) Triantis and Girod (1998), Kao and Liu (2000)).

This is not to say that other approaches such as a standard probability approach cannot address the problem of imprecision associated with the production data. Nor is it necessary to delve into the endless debate of probability versus fuzzy set theory. Others have done so extensively (see the August 1995 issue of *Technometrics* (volume 37, no. 3) and the February 1994 issue of *IEEE Transactions on Fuzzy Systems* (volume 2, no. 1) for more details). Nevertheless, one can agree with the final statement of Laviolette et al. (1995, p. 260) that states, "we maintain that

[3] Part of the material of this section is adapted from Triantis, K., Sarangi, S. and D. Kuchta, 2003, Fuzzy Pair-Wise Dominance and Fuzzy Indices: An Evaluation of Productive Performance, *European Journal of Operational Research*, 144, 412–428.

development of equitable comparative approaches is essential" and with the statement made by Laviolette and Seaman (1994. p. 38) "we have not found no evidence that fuzzy set theory is ever *exclusively* useful in solving problems."

For example, fuzzy set theory has provided an alternative modeling avenue to stochastic DEA (Sengupta 1987; Land et al. 1993; Olesen and Petersen 1995). While both of these modeling approaches can be used to address the measurement imprecision associated with the input and output variables, the data requirements are quite different. Stochastic DEA and specifically the chance-constrained programming methodology requires that the analyst "supply information on expected values of all variables, variance–covariance matrices for all variables, and probability levels at which feasibility constraints are to be satisfied" (Fried et al. 1993, p. 35).

On the other hand, the fuzzy mathematical programming approach requires the definition and measurement of the most and least plausible production bounds for all fuzzy input/output variables and an assumption on the form of their membership functions Triantis and Girod (1998). Furthermore, Almond (1995, p. 268) points outs that "being precise requires more data, and it requires more work."

It should be noted that it is also more expensive for organizations to obtain more precise data. In context of the manufacturing technology used as an illustration of the fuzzy DEA approach, Girod and Triantis (1999) were faced with the following constraint. Data necessary for the fuzzy approach, stochastic approach, and the crisp DEA approach could be obtained at a cost ranging in the hundreds, tens of thousands, and millions of dollars, respectively. This fact provides strong impetus for developing cost-effective methodologies to capture imprecision in productivity and efficiency analysis in general.

Depending on the nature of the imprecision, fuzzy numbers can represent coefficients in the objective function and/or set of constraints. Furthermore, uncertainty may also appear in the formulation of the constraints and the objective function, i.e., the extent to which the objective function is optimized and the constraints hold. During the decision-making process in a fuzzy environment, fuzzy objectives, fuzzy constraints and fuzzy decisions, represented through fuzzy sets with corresponding membership functions, can be considered. In this case, the decision is simply defined as a selection of the sets of feasible solutions or merely one solution, which simultaneously satisfies the fuzzy objective and fuzzy constraints.

However, not all data are imprecise. Some data are known with precision whereas other data are approximately known or are described with linguistic information. Methodologies need to take into account all types of available data, i.e., crisp, approximately known, and/or data expressed in a linguistic form.

There are many examples where DEA mathematical programming formulations were expanded using fuzzy set theory to incorporate the following: imprecision in the data as in Triantis and Girod (1998) and Girod and Triantis (1999), missing data and imprecise linguistic data as in Kao and Liu (2000), or fuzziness in the objective function and constraints as in Sengupta (1992), Sheth (1999), Kabnurkar (2001). Furthermore, Hougaard (1999) extends DEA technical efficiency scores

to efficiency scores defined over fuzzy intervals. The fuzzy scores allow the decision-maker to use technical efficiency scores in combination with other sources of available performance information such as expert opinions, key figures, etc.

14.2.5 Dynamical Production Systems[4]

One of the important characteristics of production and engineering systems is that they are dynamic. This characteristic needs to be captured in efficiency analyses. Samuelson (1947, p. 314) states a "system is dynamical if its behavior over time is determined by functional equations in which 'variables at different points of time' are involved in an 'essential' way."

Dynamical systems can be classified in three distinct ways: (1) dynamic and historical (Samuelson (1947)); (2) dynamic and causal (Samuelson (1947)); and (3) dynamic, causal, and closed (Vaneman and Triantis (2003)). Systems that are dynamic and historical exhibit a high degree of correlation between the variables at the initial time t_0 with the variables at the final time t. Neither the passage of time nor the structure of the system is considered, thus variables that become active during the interval between the initial time and final time, (t_0, t), are not considered when determining the final state of the system.

Dynamic and historical systems correlate the initial conditions of the system to the final conditions, and do not contain information about the structure or behavior of the system. Dynamic and historical systems can be expressed (Samuelson (1947)) as

$$y_{jt} = f\{t_0; t; x_{it_0}\} \tag{14.1}$$

where x_{it_0} is the ith input at the initial time t_0, $i = 1, 2, 3, \ldots n$ y_{jt} is the jth output at time t, $j = 1, 2, 3, \ldots m$.

$$y_{jt} = f_o; x_{it_o}$$

Dynamic and causal systems consider the initial system inputs, x_{it_0}, along with the passage of time. Dynamic and causal systems can be expressed as

$$y_{jt} = f\{t - t_0; x_{it_0}; x_{it_d}\}. \tag{14.2}$$

[4] Part of the material in this section is adapted from Vaneman, W. and K. Triantis, 2003, The Dynamic Production Axioms and System Dynamics Behaviors: The Foundation for Future Integration, *Journal of Productivity Analysis*, 19 (1), 93–113 and Vaneman and Triantis, 2007, Evaluating the Productive Efficiency of Dynamical Systems, *IEEE Transactions on Engineering Management*, 54 (3), 600–612.

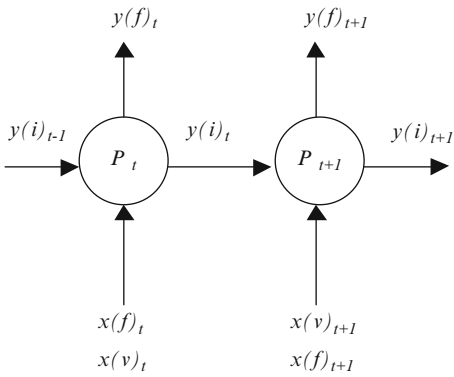


Fig. 14.1 Dynamic data envelopment analysis approach (Färe and Grosskopf 1996)

This type of system allows inputs to be added at some intermediate time t_d, where $t_0 < t_d < t$, and for the behavior of the system to be evaluated at any given time. However, the inputs added during the interval (t_0, t) are not a result of feedback mechanism from within the system.

In many applications, the systems discussed could be classified as open systems. An open system converts inputs into outputs, but the outputs are isolated from the system so that they cannot influence the inputs further. Conversely, outputs are not isolated in a closed system, thus are aware of and can influence the system behavior by providing information to modify the inputs via a feedback mechanism.

Dynamic, causal, and closed systems can be expressed as

$$y_{jt} = f\{t - t_0; x_{it_0}; x_{it_d}; y_{j(t_d - t_0)}\}, \tag{14.3}$$

where $y_{j(t_d - t_0)}$ is the jth output resulting from action during the interval $[t_0, t_d]$. By adding the output variable $y_{j(t_d - t_0)}$ to (14.3), this relationship is defined in an "essential way" such that the system brings results from past actions to influence or control future actions via a feedback mechanism

In the efficiency literature, Färe and Grosskopf (1996) were the first to study dynamical and historical systems. They defined and developed the dynamic data envelopment analysis (DDEA) by adding the element of time to the DEA model. This is accomplished by extending the static DEA model into an infinite sequence of static equations Färe and Grosskopf (1996). While this methodology does evaluate organizational performance over time, it is important to understand that this methodology is static at each discrete point in time, and provides a linear solution.

Figure 14.1 (adapted from Färe and Grosskopf 1996) illustrates the concept of DDEA. Each production cycle (P) has three types of inputs: fixed ($x(f)$), variable ($x(v)$), and inputs that were intermediate outputs to the last production cycle ($y(i)$). The outputs for each production cycle are the final outputs ($y(f)$) that represent the products that go to the customer, and intermediate outputs that are used as inputs to subsequent production cycles ($y(i)$).

As systems become more complex and time dependent, alternative methods for evaluating production and engineering system performance in a dynamic environment must be explored. The efficiency literature to date has primarily been interested in system performance measurement rather than causation, because it was thought that (1) uncovering the pattern of efficient and inefficient practices should be paramount and (2) the comparative advantage is with performance measurement and not determining the causal factors associated with system performance (Färe et al. 1994).

Nevertheless, Färe and Grosskopf (1996) also suggest the need to explore what is inside the black box of production technologies to determine how inputs are converted into outputs, so that their efficiency could be better understood. To this end, as stated in Sect. 2.2, they develop a network technology model. In their model, they evaluate how multiple inputs placed in the production process, at multiple time periods, can produce multiple outputs. While this approach is evolutionary, it fails to consider the causal relationships that exist within the network.

System performance is inherent within the system's structure and policies.[5] Thus, if the system structure (that includes physical processes with a representation of input and outputs, decisions, and organizational structure to incorporate the dissemination of information) is understood, the sources of good system performance can be replicated for future system design, and the causes of poor system performance can be corrected. Since policies are deep-seated within the system's structure, determining and evaluating causal relationships within a system will provide an understanding as to how system policies affect system performance.

To study the causal relationships, the impact of system structure and of policies, Vaneman (2002) uses system dynamics (Forrester 1961; Sterman 2000) to explore productive efficiency. System dynamics is a mathematical approach that employs a series of interrelated equations to define the system that is dynamic, closed, and causal and considers time over a continuous spectrum. System dynamics differs from traditional methods of studying productive efficiency in two fundamental ways. First, the system dynamics approach allows for problems with both dynamic and combinatorial complexity to be studied concurrently, within the same framework. Second, system dynamics is more concerned with the behavior of the system over time versus the measure of central tendency and historic trends. Differences between the DDEA approach (Färe and Grosskopf 1996) and efficiency evaluation using systems dynamics are summarized in Table 14.1.

[5] The term policy is used to describe how a decision process converts information into action to show a change in the system (Forrester (1961)). Forrester (1968) further identifies four concepts found within any policy statement:
1. A goal.
2. An observed condition of the system.
3. A method to express any discrepancy between the goal and the observed condition.
4. Guidelines of which actions to take based on the discrepancy.

Table 14.1 Differences between the DDEA (Färe and Grosskopf 1996) and system dynamics approaches when measuring productive efficiency (adapted from Vaneman 2002)

Attribute	Dynamic data envelopment analysis (DDEA)	System dynamics approach for measuring productive efficiency
Primary authors	Färe and Grosskopf (1996)	Coyle (1996), Wolstenholme (1990), Vaneman (2002)
Goal	Decision-making based upon efficiency measurement, and estimation of the effects of policy change over time	Decision-making based upon the behavior of the endogenous elements of the system with respect to efficiency measurement, and the simulated effects of policy changes over time
Technical Approach	Optimization through linear programming	Optimization through system dynamics (Heuristic approach)
Characteristics	Dynamic	Dynamic
	Deterministic or probabilistic	Deterministic or probabilistic
	Linear	Linear or non-linear
	Discrete time	Continuous time
Advantages	Linear programming approach guaranteed to find optimal solution	Model represents causal relationships well
	Optimal solution is easily interpreted	Suggested policy changes are calculated and simulated over time
		Model allows for nonlinear relationships
		Model allows for feedback within the structure
		Model can represent information flows
Disadvantages	Model does not allow for causal relationships to be defined	Heuristic approach not guaranteed to find optimal solution
	Policy changes and their effects are estimated (not simulated) and observed over time	Optimal solution is not always readily apparent
	The model only accommodates linear relationships	
	Does not allow for feedback within networks	
	Does not allow for information flows to be modeled	

Vaneman and Triantis (2007) expand on a methodological approach that combines the system dynamics (SD) modeling paradigm with concepts taken from the measurement of productive efficiency so as to evaluate dynamical systems. In this research, the SD paradigm is coupled with the fundamental assumptions of production theory in order to evaluate productive efficiency performance. As a result, a structure within the SD model is introduced that computes efficiency performance scores based on a hill-climbing optimization procedure. The structure is illustrated using an electric utility example described by Kopp (1981) where constrained optimization results are not only replicated but the path to achieve optimal dynamic system performance is also found. An example is provided of

how this structure can be used to facilitate policy decisions within a technology management environment where investments are made in new technologies.

Furthermore, in the context of evaluating the efficiency performance of dynamic, causal, and closed systems, one must also investigate their equilibrium and stability. Systems can be categorized as being in equilibrium or in disequilibrium, stable or unstable. Equilibrium can be further categorized as either static or dynamic. Static equilibrium is defined as the condition that exists when there is no flow within the system (Sterman 2000). Two conditions must be satisfied for a system to be in static equilibrium: (1) all first order derivatives x'_{it}, y'_{jt} are zero at the time considered and (2) all higher order derivatives are also zero. A system in which only condition (1) is satisfied is said to be momentarily at rest (Frisch 1935–1936).

A system in dynamic equilibrium is a system where there is a constant flow going through the system. Viewing the system from a macrolevel, dynamic equilibrium gives the appearance that nothing within the system changes over time. A closer look reveals that there is a constant flow of inputs into the system, and a constant flow of outputs from the system (Sterman 2000). All derivatives will have nonzero values for dynamic equilibrium.

System stability refers to how a system that was previously in equilibrium behaves when a disturbance is introduced. Consider a small disturbance introduced to the system at time t_d. If the system returns to its original (or closely related) state of equilibrium after being disturbed, the system is considered stable (Frisch 1935–1936; Sterman 2000). If the small disturbance forces the system further away from equilibrium with the passage of time, the system is said to be in unstable equilibrium (Sterman 2000).

14.2.6 Visualization of the DEA Results: Influential Data Identification

One of the important aspects of making effective decisions for engineering applications is the need to improve upon the existing visual representation of the DEA results. The question that needs to be addressed and answered is how to incorporate in DEA existing computer science, statistical, and graphical-based visualization techniques so that decision-makers can more effectively use the information generated by the DEA analytical framework. These results primarily include the DEA scores, the peers, and the performance targets.

However, any discussion on data visualization inevitably leads to a subsequent discussion on outliers and influential observations. This is because what become apparent from any visualization approach are the extremes in efficiency performance, i.e., the efficient and extremely inefficient observations along with the main mass of observations.

In general, influential data points fall into three categories: (1) observations that are outliers in the space of the independent variables (referred to as leverage

influence). (2) Points that deviate from the linear pattern of the majority of the data (referred to as residual influence). (3) Points that have both leverage and residual influence (referred to as interaction influence).

There are three main issues that arise from the research with respect to influential observation in efficiency analysis. The first has to do with identification, i.e., how to identify the influential observations given the masking effect that exists in most datasets (Seaver and Triantis 1989, 1992, 1995; Wilson 1993, 1995). The second issue has to do with the impact that each influential observation has on the efficiency scores. The third and often misguided issue is what to do with influential observations once they are identified. The temptation is to remove extreme observations from the data set. However, in most cases extreme observation are the observations that have the most information as they represent extreme production occurrences. Therefore, the identification of influential observations allows the decision-maker and engineer to isolate important observations that represent unique occurrences that can be subsequently used for benchmarking purposes and process improvement interventions.

Nevertheless, irrespective of the methods used for outlier and influential observation identification and the assessment of their impact on the efficiency scores, there is a need to study existing data visualization approaches and their appropriateness in effectively representing DEA results to decision-makers.

14.3 A DEA-Based Approach Used for the Design of an Integrated Performance Measurement System

Based on the implementation of DEA in manufacturing (Triantis and Girod 1998; Girod and Triantis 1999) and service environments (Triantis and Medina-Borja 1996) as part of the funded research activities of the System Performance Laboratory at Virginia Tech, an approach has been developed that allows the researcher not only to define appropriate efficiency-based performance measures but also to link these performance measures to performance improvement actions.

Initially, one needs to communicate to decision-makers that organizational performance is difficult to measure and improve. The complexity of organizations no longer allows performance measurement to be addressed with "off-the-cuff" solutions. Furthermore, one needs a language to represent production complexity, i.e., to understand how the production structure of the organization creates the behavior that is observed. One can represent this organizational production complexity with multiple key variables that directly or indirectly account for the fashion in which inputs are transformed into outputs. Additionally, it is necessary to evaluate the organization's current key performance measures or indicators in terms of how they account for the different organizational performance dimensions such as cost, quality, availability, safety, capacity utilization, etc.

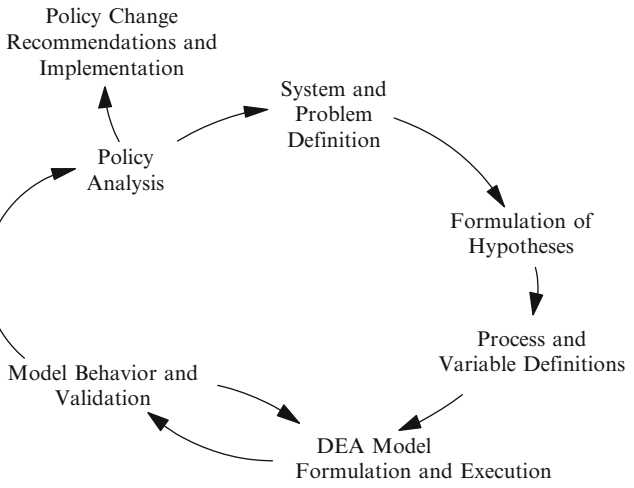


Fig. 14.2 A DEA-based approach used to develop an integrated performance measurement system

This leads to the requirement to design and implement a performance evaluation system that is based on DEA. This system should integrate existing key performance indicators or measures, key variables that represent the organization's production structure as well as the data used to represent both the measures and variables. The integrated performance measurement system should provide performance benchmarks, identify best practices, define performance targets, and provide input to the operational planning process of the organization that will plan for process and organizational performance improvements.

The proposed system can be used to address key questions such as:

1. Given the plethora of performance measures collected by the organization, which one should management take into account when focusing on key performance dimensions (such as cost, quality, etc.)?
2. How should the organization integrate its key performance measures given the fact that different sources of data are currently used?
3. Are current performance measurement systems effective in terms of meeting its current goals and affecting change?
4. How do lessons learned from the performance improvement interventions feedback into the definition of performance objectives?
5. How does one make a fair comparison among different DMUs within the organization using the key performance measures?
6. How does one identify achievable performance improvement interventions for each DMU?

Figure 14.2 provides a DEA-based approach that can be used to design the integrated performance measurement system.

1. *Task 1*: The first task involves two critical activities. First, a measurement team should be established. The team will include representatives from the organization and the researchers/analysts. The measurement team should interact in team meetings with the decision-makers and should provide key information for the analytical model. Second, the performance measurement problem should be defined and the purpose and use of the model should be established.
2. *Task 2*: Once the problem is identified in Task 1, theories about the causes of thecurrent organizational performance levels can be developed. These theories, also known as hypotheses, are believed to account for the behavior demonstrated by the organization. The fundamental hypothesis is that the variables designated as inputs determine the response of the output variables that will be selected to represent the behavior of the organization. Additionally, it will be hypothesized that specific infrastructure and environmental variables also impact performance. Once the most appropriate variables representing the behavior of the organization are selected, their units of measurement should be defined, and an assessment will be made as to how to best obtain data for these variables.
3. *Task 3*: The third task is the system conceptualization. This task constitutes the beginning of model development. At this point there are a number of activities that are required. First, the input and output variables selected in Task two will require further classification. These variables will be classified as representing good or bad inputs/outputs, whether they are controllable (discretionary) or not, whether they represent quality or quantity performance, whether they are fixed or variable, whether they represent environmental conditions or not, whether they represent infrastructure etc. Second, uncertainty or imprecision associated with the measurement of specific variables should be determined. Third, various appropriate input/output specifications should be identified. This implies that there may be more than one input/output model that is appropriate for the evaluation of the organization.
4. *Task 4*: The DEA model formulation is the next task. During this task, the DEA input/output specifications generated during the system conceptualization phase are transformed into mathematical models so they can be "solved." At this point, solution algorithms will be proposed and the DEA formulations will be programmed in the appropriate software platform.
5. *Task 5*: The DEA model behavior and validation task will be conducted during this task under the premise that all models of the real world are inaccurate to some extent. These inaccuracies occur because it is difficult to model every variable that affects the performance of the organization. The results of the DEA formulations will be evaluated considering the identification of the best and worst performing DMUs, the benchmark DMUs, and the performance targets.
6. *Task 6:* At this point, the policy analysis can begin. The DEA results will point to different operational interventions that decision-makers can make.
7. *Task 7*: The last task is to plan for the implementation and deployment of the policies for a real world system. By this time policies have been evaluated and tested under numerous conditions to ensure that they represent the direction best for the real world system.

The first six tasks outlined in this section will be completed in an iterative fashion. The first iteration will require extraction of a very small number, but important, factors (variables) associated with the performance of the organization, collect data, formulate and solve the model, test and validate it, and then evaluate alternative performance improvement interventions. During the second iteration additional factors (variables) will be added and the first six tasks will be repeated. The process will be repeated until a point is reached where all the important elements of the problem are included. Finally, completing task seven will finalize the process.

14.4 Selected DEA Engineering Applications

This section provides an overview of some of the existing work on DEA engineering applications. Not all applications are described in this section. The remaining applications are found in the bibliography at the end of this chapter.

In the first part of this section (Sect. 4.1) the four applications presented at the 1999 NSF Workshop on Engineering Applications of DEA at Union College are summarized. In the remaining parts of this section, existing applications that consider the role of environmental issues in production (Sect. 4.2), that evaluate the performance of transportation systems (Sect. 4.3), and that present studies in other areas of engineering (Sect. 4.4) are briefly presented.

14.4.1 Four Applications

14.4.1.1 Evaluating Efficiency of Turbofan Jet Engines (Bulla et al. 2000)

DEA is explained and illustrated with an application to data on turbofan jet engines. The purpose of this study is to augment, not replace, traditional methods for measuring efficiency. What is interesting is that the results of this study are compared with standard engineering methods for measuring efficiency. This is something that more researchers should focus on in the future. It should be noted that the input and output variables are defined based on the turbofan engine characteristics. In contrast to the engineering methods for measuring efficiency, DEA handles data on multiple inputs (fuel consumption, engine weight, drag) and multiple outputs (airflow and cruise thrust) that can be used to characterize the performance of each engine. Furthermore, the DEA analysis is accomplished without any use of prearranged engineering-based weights. DEA also provides information on the sources and amounts of inefficiencies in each input and output of each engine without requiring knowledge of the functional relations among the inputs and outputs.

14.4.1.2 Measurement and Monitoring of Relative Efficiency of Highway Maintenance Patrols (Cook et al. 1990, 1994) and of Highway Maintenance Operations (Ozbek et al. 2010a, b)[6]

This study was motivated in part from the fact that the Ontario Ministry of Transportation in Canada needed a management tool by which to assess quantitatively the relative efficiency of the different patrols. Additionally, a desire was voiced to evaluate the impact of different "external" (regional) factors as well as maintenance policy options (e.g., extent on privatization) on patrol efficiency. Comparisons were made both region wide and province wide.

The innovation of this research stems from the definition of inputs (maintenance expenditure, capital expenditure, and climate factor) and outputs (assignment size factor, average traffic served, rating change factor, and accident prevention factor). More specifically, the area served factor captures the extent of the workload for which the patrol has responsibility. The average traffic served is a measure of the overall benefit to the users of the highway system in a patrol. The pavement rating change factor measures the actual change in the various road sections. The accident prevention factor presents the accident prevention goal of maintenance. The maintenance and capital expenditures are straightforward. However, the climatic factor takes into account snowfall, major/minor temperature cycles, and rainfall.

The authors conclude by emphasizing the extreme importance to choose the correct input/output factors that will enter the DEA analysis. Furthermore, the unbounded version of DEA does not satisfactorily capture real-life situations. However, there is no "mechanical" process for choosing bounds on the inputs and outputs. What is required is a thorough knowledge of the process in which the DMUs are engaged, a clear vision for the purposes for which efficiency is measured, and sound managerial considerations.

Ozbek et al. (2010a, b) provide a performance measurement framework for road maintenance operations. Within the last two decades, the road maintenance concept has been gaining tremendous attention. This has brought about new institutional changes, predominant of which is the challenge for maintenance managers to achieve maximum performance from the existing road system. Such challenge makes it imperative to implement comprehensive systems that measure road maintenance performance. However, as pointed out by the Transportation Research Board in 2006, even though the road maintenance performance measurement systems developed and implemented by the state departments of transportation elaborate on the maintenance level-of-service (i.e., effectiveness of the road maintenance), the fundamental relationships between the maintenance level-of-service and the budget requirements (i.e., efficiency of road maintenance) need

[6] Part of this discussion has been taken from Ozbek et al. 2010a, b.

more investigation. This is mainly because not knowing how "efficient" state departments of transportation are in being "effective" can lead to excessive and unrealistic maintenance budget expectations.

In an effort to address this need, the research by Ozbek et al. (2010a, b) develops and implements a comprehensive framework that can measure the overall efficiency of road maintenance operations. This framework is designed to consider the effects of environmental (e.g., climate, location, etc.) and operational (e.g., traffic, load, etc.) factors on such overall efficiency. This research specifically provides an overview of the data and modeling issues faced during the early stages of the implementation of the framework to the Virginia Department of Transportation's case for the maintenance of bridges and presents the core implementation stages of the framework in trying to identify (1) the relative efficiency of seven counties of Virginia in performing bridge maintenance, (2) the benchmarks (peers) and targets that pertain to the inefficient counties, and (3) the effects of the environmental and operational factors on the road maintenance efficiency of counties.

14.4.1.3 Data Envelopment Analysis of Space and Terrestrially Based Large Scale Commercial Power Systems for Earth (Criswell and Thompson 1996)

This research addresses an important energy problem that society will face in the twenty-first century. The Lunar Solar Power system (LSP) is a new option that is independent of the biosphere. LSP captures sunlight on the moon, converts the solar power to microwaves, and beams the power to receivers on Earth the output electricity. The collimated microwave beams are low in intensity, safe, and environmentally benign. LSP is compared to fossil, fission, thermal, and nuclear technologies using DEA. The comparison suggests that the efficiencies that are gained from LSP are large. In terms of a normalized cost/benefit ratio, DEA reveals that LSP is much more efficient than the other technologies. What is noteworthy is that the gains remain even if the resources needed for LSP are tenfold greater than what is estimated from governmental studies.

14.4.1.4 The Relationship of DEA and Control Charts (Hoopes and Triantis 2001)

It is the contention of Hoopes and Triantis (2001) that measurement of efficiency performance in conjunction with statistical process control charts can be the starting point from which the causal factors can be identified and process improvement interventions can be initiated. Consequently, given this challenge, the following objective is identified for this study, i.e., to provide a conceptual linkage between efficiency performance assessment and the traditional control charts (Shewhart 1980). In order to create the conceptual linkage to traditional

control charts, the specification of the production function in this study uses the concepts of process and product characteristics. Control charts evaluate the stochastic behavior of the production process by studying process and/or product characteristics one at time. On the other hand, efficiency measurement approaches include as part of their evaluation the entire set of critical product and/or process characteristics simultaneously.

This research shows that these two approaches can be used in a complementary manner to identify unusual or extreme production instances, benchmark production occurrences, and evaluate the contribution of individual process and product characteristics to the overall performance of the production process. The identification of extreme production instances in conjunction with the evaluation of their technological and managerial characteristics is used to identify potential root causes. Decision-makers can use this information to make process improvements. These issues are illustrated by studying the inner layer production process of a circuit board manufacturing facility. The inner layer process is one of the first stages in a manufacturing system that produces multilayer printed circuit boards.

14.4.2 The Effect of Environmental Controls on Productive Efficiency[7]

The impact of pollution prevention methods on process efficiency performance has been a concern of a number of studies. This issue is extremely important for chemical process manufacturing organizations. In the literature, the generation of pollution has been taken into account using a number of different analytical methods. Two fundamental assumptions underlying these methods are that (a) pollution controls are after-the-fact, end-of-pipe and do not explicitly take into account process and other changes that can reduce pollution (i.e., pollution prevention) and can potentially improve efficiency performance. (b) Interactions with other production systems represented by variations in input and output mixes are not taken into account.

In measuring productive efficiency, waste products are sometimes explicitly considered and sometimes not. Both econometric and DEA-based techniques are used to evaluate the effect of pollution controls on productive performance. There are three approaches that have been used in the literature. One is to apply standard DEA or econometric analysis to DMUs with and without pollution controls. The results obtained by considering and not considering pollution controls provide estimates of the cost of lost production associated with pollution controls. The other two approaches explicitly consider pollution or undesirable outputs and determine

[7] Part of the material in this section is adapted from Triantis, K. and P. Otis, 2003, A Dominance Based Definition of Productive Efficiency for Manufacturing Taking into Account Pollution Prevention and Recycling, forthcoming, *European Journal of Operational Research*.

the cost of pollution control based on loss of disposability or assign shadow prices to the undesirable outputs.

Tyteca (1995, 1996) compares several methods of evaluating environmental performance of power plants. The methods are those developed by Färe et al. (1989) where both output-oriented and input-oriented efficiencies are considered and an approach developed by Hayes et al. (1993) that maximizes ratios of weighted sums of desirable outputs to weighted sums of pollutants where pollutants are treated as inputs. In addition, an index method is used to calculate performance metrics as a means of comparison. Depending on the approach chosen, significant variations were found in the relative efficiencies of the power plants evaluated.

Recently, Wang et al. (2002) propose a DEA model where desirable and undesirable outputs are treated synchronously. A sample of efficiency measures for ten paper mills is given. Ramanathan (2002) considers several variables simultaneously when comparing carbon emissions of countries. He illustrates the use of the DEA methodology with four variables that include CO_2 emissions, energy consumption, and economic activity. Soloveitchik et al. (2002) use DEA in conjunction with multiple objective optimization models to examine the long-run capacity expansion problem of power generation systems as a base for defining the marginal abatement cost. Sarkis and Weinrach (2001) analyze a decision-making case study concerning the investment and adoption of environmentally conscious waste treatment technology in a government-supported agency. Sarkis (1999) provides a DEA-based methodological framework for evaluating environmentally conscious manufacturing programs. Finally, Otis (1999) expands the Full-Disposal Hull (Deprins et al. 1984) to evaluate the impact of pollution prevention on efficiency performance of circuit board manufacturing facilities.

14.4.3 The Performance of Transit Systems

DEA has also been used to evaluate the performance of public transit systems. For example, Husain et al. (2000) present a study that evaluates the performance of services of a public transit organization in Malaysia. Chu et al. (1992) compare transit agencies operating in the USA, whereas Boile (2001) evaluates public transit agencies. Tone and Sawada (1991) evaluate bus enterprises under public management and explore the notions of service efficiency, cost efficiency, income efficiency, and public service efficiency separately. Nolan (1996) applies DEA to evaluate the efficiency of mid-sized transit agencies in the USA using the data from Section 15 US DOT 1989–1993. Carotenuto et al. (2001) stress the fact that in recent years, it has become imperative for public transit organizations to rationalize the operating costs and to improve the quality of the services offered. They obtain measures of pure technical, scale and overall efficiency of both public and private agencies. Kerstens (1996) investigates the performance of the French Urban Transit sector by evaluating the single mode bus operating companies and Dervaux et al. (1998) compute radial and nonradial measures of efficiency

and compute congestion. Nakanishi et al. (2000) estimate the relative efficiency of transit agencies providing motorbus service. Odeck (2000) assesses the relative efficiency and productivity growth of Norwegian motor vehicle inspection services for the period 1989–1991 using DEA and Malmquist indices. Finally, Cowie and Asenova (1999) examine the British bus industry in light of fundamental reform in ownership and regulation.

As seen by the utilization of DEA for the purpose of evaluating transit systems, the research primarily focuses of the assessment of companies and/or agencies. Very little research has focused on the evaluation of performance at the transportation process level. Notable exceptions include the evaluation of two-way intersections by Kumar (2002), the prioritization of highway accident sites by Cook et al. (2001), and bus routes by Sheth et al. (2007).

14.4.4 Other Engineering Applications of DEA

Some of the most recent engineering applications of DEA focus on two primary issues, i.e., evaluating the investment in equipment and assessing alternative organizational structures (such as worker teams).

Sun (2002) reports on an application of DEA to evaluate computer numerical control (CNC) machines in terms of system specification and cost. The evaluation contributed to a study of advanced manufacturing system investments was carried out in 2000 by the Taiwanese Combined Service Forces. The methodology proposed for the evaluation of the 21 CNC machines is based on the combination of the Banker, Charnes, and Cooper (BCC) model (Banker et al. 1984) and cross-efficiency evaluation (Doyle and Green 1991). Both Karsak (1999) and Braglia and Petroni (1999a, b) use DEA to select industrial robots. Karsak (1999) argues that a robust robot selection procedure necessitates the consideration of both quantitative criteria such as cost and engineering attributes, and qualitative criteria, e.g., vendor-related attributes, in the decision process. The qualitative attributes are modeled using linguistic variables represented by fuzzy numbers. Braglia and Petroni (1999a, b) adopt a methodology that is based on a sequential dual use of DEA with restricted weights. This approach increases the discriminatory power of standard DEA and makes it possible to achieve a better evaluation of robot performance. Finally, Sarkis and Talluri (1999) present a model for evaluating alternative Flexible Manufacturing Systems by considering both quantitative and qualitative factors. The evaluation process utilizes DEA model, which incorporates both ordinal and cardinal measures.

Paradi et al. (2002) propose DEA as an approach to evaluate knowledge worker productivity. DEA is used to examine the productivity, efficiency, and effectiveness of one such knowledge worker group – the Engineering Design Teams (EDT) at Bell Canada, the largest telecommunications carrier in Canada. Two functional models of the EDT's were developed and analyzed using input-oriented constant returns to scale (CRS) and variable returns to scale (VRS) DEA models.

Finally, Miyashita and Yamakawa (2002) study collaborative product design teams. This research stems from the need to reduce the time for product development where engineers in each discipline have to develop and improve their objectives collaboratively. Sometimes, they have to cooperate with those who have no knowledge at all for their own discipline. Collaborative design teams are proposed to solve these kinds of the problems and consequently their effectiveness needs to be assessed.

14.5 Systems Thinking Concepts and Future DEA Research in Engineering

One of the skills that engineers are taught and practice is that of systems thinking and modeling (O'Connor and McDermott 1997; Richmond 2000; Sterman 2000). Furthermore, systems engineering as a discipline is gaining more and more acceptance. Growth in this discipline is expected in the future as it brings together not only all aspects of engineering but also the management of engineering practices, functions, and processes. Perhaps, the future success of implementing the DEA approach in engineering is to communicate through a systems-based framework in terms of defining the performance measurement problem, stating and testing hypotheses, and implementing performance improvement strategies. This framework would not be inconsistent with the firm/organizational/process view that the DEA research has held over the years. Nevertheless, there are three, if not more, systems-related issues briefly described in this section that need to be considered when completing DEA research in engineering. In fact, future researchers should explore other systems thinking issues that are pertinent for DEA research.

14.5.1 The Need for Operational Thinking

To open the "input/output transformation box" as discussed in Sect. 2.2, the researcher needs to understand the operational structure of the production process that is being represented by the DEA model. Operational thinking addresses the question about how performance is actually being generated. One can say that operational thinking is concerned with the "physics" of how systems use scarce resources to add value and generate outputs. The researcher attempts to answer the question: "How does this actually work?"

Operational thinking as indicated in Sect. 2.2 requires access to and investment in time with engineering/manufacturing organizations. The difficulty in getting access to these organizations is that they do not want to necessarily invest time in researchers who would like to "learn" about their processes. Decision-makers do

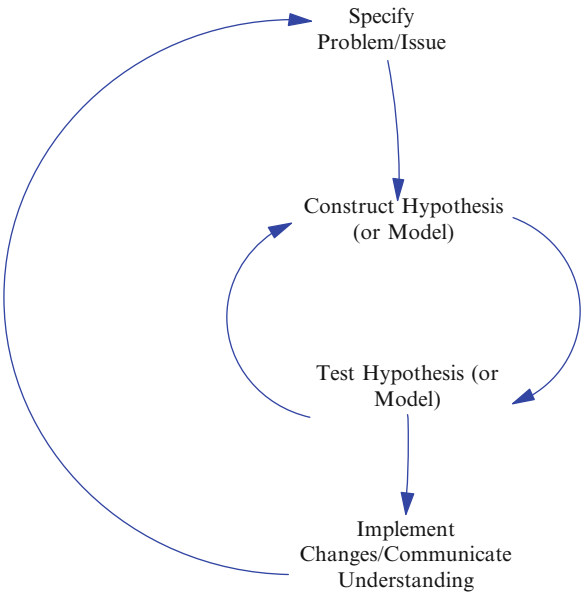


Fig. 14.3 Contribution of modeling to the scientific process

not immediately see the long-term benefits that may accrue from the analytical production studies. However, both organizations and the researchers have to meet half way. The researchers need to be willing to make problem-solving contributions to the organizations during their visits within organizations and decision-makers need to learn about and appreciate the benefits that they will accrue from the analytical studies.

14.5.2 Contribution to Performance Measurement Science

It is perhaps not stated often enough that one of the fundamental contributions of DEA research is its impact on the science of performance measurement and improvement. As such, a discussion needs to be initiated by the research community that will further define the scope of this scientific domain and will more specifically identify the impacts of the DEA modeling framework.

Fundamentally, as Fig. 14.3 indicates, the researcher starts with a problem (perhaps some type of performance behavior), formulates a theory/hypothesis that is primarily represented by the DEA model and to the extent that the model can generate the problem (the observed performance behavior), the researcher has a reasonable hypothesis/theory and model. If not, one must modify the model and retest it and/or modify the hypothesis/theory. Once the researcher has arrived at a reasonable hypothesis/theory, one can then communicate the newfound clarity to others and begin to implement process changes.

Scientific thinking is most applicable *after* one has constructed the model since it helps to build a better, shared understanding of a system that in our case is a production process. Furthermore, discarding current paradigms marks progress in science. However, it is not usual in DEA research to emphasize counterintuitive results that challenge current paradigms.

Nevertheless, the term "model" represents our assumptions (or hypotheses) about how a particular part of the world works. The value in modeling rests in how *useful* the model is in terms of shedding light on an issue or problem. As Deming among others has stated that all models are wrong and some are useful. What has been missing from the DEA literature is an assessment of how useful specific DEA models have been.

Nevertheless, system thinkers focus on choosing numbers that are simple and easy to understand and that make sense relative to one another. They justify doing this because they believe that insight arises from understanding the *relationships* between numbers rather than the absolute numbers themselves. This is a fundamental notion in systems thinking, which is consistent with the concept of relative efficiency performance in DEA.

Furthermore, seasoned system thinkers continually resist the pressure to "validate" their models (i.e., prove truth) by tracking history. Instead, they work hard to become aware of where their models cease to be useful for guiding decision-making and to communicate their models to the decision-makers. This means that they are mostly concerned with face-validity (model structure and robustness). Face-validity tests that assess how well the structure of a system model matches the structure of the reality of the model is intended to represent (Richmond 2000). For example, do the input/output variables truly represent the underlying production technology?

Systems thinkers also pay a lot of attention to model robustness, i.e., they torture-test their models. They want to know under what circumstances their models "break down." They also want to know, "Does it break down in a realistic fashion." There is a whole set of variables in a production environment whose behavior patterns and interrelationships show an internal consistency. To judge robustness of the model, one asks whether its variables show the same kinds of internal consistency (Richmond 2000). These issues of model structure and robustness need to be addressed further by DEA researchers in the context of assessing the scientific contributions of their DEA research.

14.5.3 Relationship of the DEA Model with the Real World

During the DEA research process, the relationship between the real and virtual world represented by the DEA model should be constantly kept in mind. A DEA modeling effort should have the primary goal of establishing policies or procedures that can be implemented in the real world. However, many modelers and performance measurement teams often lose sight of the real world implementation.

In the virtual world the system structure is known and controllable. Information is complete, accurate and the feedback is almost instantaneous from actions within the system. Decisions and policy implementations are usually perfect. The goal of the virtual world is learning (Sterman 2000). One interesting question to investigate within the DEA research community is the extent of learning that is achieved by the DEA modeling process.

The goal of real world systems is performance. In the real world, systems often have unknown structures and experiments are uncontrollable. Given these factors, policies that are implemented in the real world system first often do not realize their full impact until long after the policy is implemented. Information from real world systems is often incomplete due to long delays, and data are often biased or contain errors (Sterman 2000).

The erroneous or incomplete data often lead to inaccurate mental models. Mental models serve as the basis for one's beliefs of how a system works. Receiving inaccurate or incomplete data about the system often leads to erroneous beliefs of how the system performs. This condition leads to the establishments of rules, system structure, or strategies that govern the system in a suboptimal manner. When strategies, policies, structure, or decision rules are inaccurate, decisions made within the system are also incorrect. This creates negative consequences for the system processes and leads to a vicious circle for declining system performance.

However, DEA modeling should offer the decision-maker an opportunity to learn about system behavior, ask pertinent questions, collect accurate information, and define more effective policies. The constant and effective interface between the virtual and real world allows for better decision-making over the long run.

References

Adolphson D, Cornia G, Walters L. "A unified framework for classifying DEA models. In: Bradley H, editor. Operational research 90. Oxford, UK: Pergamon Press; 1990. p. 647–57.

Aigner DJ, Lovell CAK, Schmidt P. Formulation and estimation of stochastic frontier production functions. J Econ. 1977;6(1):21–37.

Akiyama T, Shao CF. Fuzzy mathematical programming for traffic safety planning on an urban expressway. Transp Plan Technol. 1993;17:179–90.

Akosa G, Franceys R, Barker P, Weyman-Jones T. Efficiency of water supply and sanitation projects in Ghana. J Infrastruct Syst. 1995;1(1):56–65.

Al-Majed M. 1998, Priority-Rating of Public Maintenance Work in Saudi Arabia, M.S. Thesis, King Fahd University of Petroleum and Minerals, Saudi Arabia.

Almond RG. Discussion: fuzzy logic: better science or better engineering? Technometrics. 1995;37(3):267–70.

Amin G, Shirvani MS. Evaluation of scheduling solutions in parallel processing using DEA/FDH model. J Ind Eng Int. 2009;5(9):58–62.

Amos J. Transformation to agility (manufacturing, aerospace industry. Ph.D. dissertation, The University of Texas at Austin, 1996.

Anandalingam G. A mathematical programming model of decentralized multi-level systems. J Oper Res Soc. 1988;39:1021–33.

Anastasopoulos PC, McCullouch BG, Gkritza K, Mannering FL, Kumares SC. A Cost Saving Analysis for Performance-based Contracts For Highway Maintenance Operations. ASCE J Infrastruct Syst 2009. http://dx.doi.org/10.1061/(ASCE)IS.1943-555X.0000012.

Anderson T, Hollingsworth K. An introduction to data envelopment analysis in technology management. In: Kacaoglu D, Anderson T, editors. Portland conference on management of engineering and technology. New York: IEEE; 1997. p. 773–8.

Athanassopoulos A. Goal programming and data envelopment analysis (GoDEA) for target-based multi-level planning: allocating central grants to the Greek local authorities. Eur J Oper Res. 1995;87:535–50.

Athanassopoulos A, Lambroukos N, Seiford L. Data envelopment scenario analysis for setting targets to electricity generating plants. Eur J Oper Res. 1999;115(3):413–28.

Bagdadioglu N, Price C, Weymanjones T. Efficiency and ownership in electricity distribution-a nonparametric model of the Turkish experience. Energ Econ. 1996;18(1/2):1–23.

Baker R, Talluri S. A closer look at the Use of data envelopment analysis for technology selection. Comput Ind Eng. 1997;32(1):101–8.

Wang CH. 1993, The Impact of Manufacturing Performance on Firm Performance, the Determinants of Manufacturing Performance and the Shift of the Manufacturing Efficiency Frontier, Ph.D. Dissertation, State University of New York in Buffalo.

Banker RD, Datar SM, Kermerer CF. A model to evaluate variables impacting the productivity of software maintenance projects. Manag Sci. 1991;37(1):1–18.

Banker RD, Kermerer CF. Scale economies in new software development. IEEE Trans Softw Eng. 1989;15(10):1199–205.

Banker RD, Charnes A, Cooper WW. Some models for estimating technical and scale efficiencies in data envelopment analysis. Manag Sci. 1984;30(9):1078–92.

Bannister G, Stolp C. Regional concentration and efficiency in Mexican manufacturing. Eur J Oper Res. 1995;80(3):672–90.

Battese GS, Corra GS. Estimation of a production frontier model: with application to the pastoral zone of eastern Australia. Aust J Agric Econ. 1977;21:169–79.

Bellman RE, Zadeh LA. Decision-making in a fuzzy environment. Manag Sci. 1970;17 (4):141–64.

Boggs RL. Hazardous waste treatment facilities: modeling production with pollution as both an input and output. The University of North Carolina at Chapel Hill: Ph.D. Dissertation; 1997.

Boile MP. Estimating technical and scale inefficiencies of public transit systems. J Transp Eng. 2001;127(3):187–94. ASCE.

Bookbinder JH, Qu WW. Comparing the performance of major American railroads. J Transport Res Forum. 1993;33(1):70–83.

Borger B, Kerstens K, Staat M. Transit cost and cost efficiency: bootstrapping non-parametric frontiers. Res Transp Econ. 2008;23:53–64.

Borja A. Outcome based measurement of social service organizations: a DEA approach, Dissertation, Virginia Tech, Department of Industrial and Systems Engineering, Falls Church, VA: Ph.D; 2002.

Bowen WM. The nuclear waste site selection decision-A comparison of Two decision-aiding models. Ph.D: Dissertation, Indiana University; 1990.

Bowlin WF. Evaluating the efficiency of US Air force real-property maintenance activities. J Oper Res Soc. 1987;38(2):127–35.

Bowlin WF, Charnes A, Cooper WW. Efficiency and effectiveness in DEA: an illustrative application to base maintenance activities in the US air force. In: Davis OA, editor. Papers in cost benefit analysis. Pittsburgh, PA: Carnegie-Mellon University; 1988.

Braglia M, Petroni A. Data envelopment analysis for dispatching rule selection. Prod Plann Contr. 1999a;10(5):454–61.

Braglia M, Petroni A. Evaluating and selecting investments in industrial robots. Int J Prod Res. 1999b;37(18):4157–78.

Bulla S, Cooper WW, Wilson D, Park KS. Evaluating efficiencies of turbofan Jet engines: a data envelopment analysis approach. J Propul Power. 2000;16(3):431–9.
Busby JS, Williams GM, Williamson A. The Use of frontier analysis for goal setting in managing engineering design. J Eng Des. 1997;8(1):53–74.
Byrnes P, Färe R, Grosskopf S. Measuring productive efficiency: an application to Illinois strip mines. Manag Sci. 1984;30(6):671–81.
Campbell DG, Frontiers P. Technical efficiency and productivity measurement in a panel of united states manufacturing plants. Ph.D: Dissertation, University of Maryland; 1993.
Caporaletti L, Gillenwater E. 1995, The Use of Data Envelopment Analysis for the Evaluation of a Multiple Quality Characteristic Manufacturing Process, *37th Annual Meeting-Southwest Academy of Management,* C. Boyd, ed., 214–218, Southwest Academy of Management.
Carbone TA. 2000, Measuring efficiency of semiconductor manufacturing operations using Data Envelopment Analysis (DEA), Proceedings of IEEE International Symposium on Semiconductor Manufacturing Conference, IEEE, Piscataway, NJ, USA, 56–62.
Cardillo D, Tiziana F. DEA model for the efficiency evaluation of non-dominated paths on a road network. Eur J Oper Res. 2000;121(3):549–58.
Carotenuto, P., Coffari A., Gastaldi, 1997, M., and N. Levialdi, Analyzing Transportation Public Agencies Performance Using Data Envelopment Analysis, *Transportation Systems IFAC IFIP IFORS Symposium*, Papageorgiou, M. and A. Poulieszos, editors, 655–660, Elsevier.
Carotenuto, P. Mancuso, P. and L. Tagliente, 2001, Public Transportation Agencies Performance: An Evaluation Approach Based on Data Envelopment Analysis, NECTAR Conference No 6 European Strategies in the Globalizing Markets; Transport Innovations, Competitiveness and Sustainability in the Information Age, 16–18 May, Espoo Finland.
Celebi, D. and D. Bayraktar, 2008, An Integrated Neural Network and Data Envelopment Analysis for Supplier Evaluation under Incomplete Information, *Expert Systems with Applications*, 1648–1710.
Chai DK, Ho DC. Multiple criteria decision model for resource allocation: a case study in an electric utility. Infor. 1998;36(3):151–60.
Chang K-P, Kao P-H. The relative efficiency of public-versus private municipal bus firms: an application of data envelopment analysis. J Product Anal. 1992;3:67–84.
Chang YL, Sueyoshi T, Sullivan RS. Ranking dispatching rules by data envelopment analysis in a Job shop environment. IIE Trans. 1996;28(8):631–42.
Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision-making units. Eur J Oper Res. 1978;2:429–44.
Charnes A, Cooper WW, Lewin A, Seiford L, editors. Data envelopment analysis: theory, methodology and applications. Norwell, MA: Kluwer; 1994.
Chen T-Y. An assessment of technical efficiency and cross-efficiency in Taiwan's electricity distribution sector. Eur J Oper Res. 2002;137(2):421–33.
Chen W. The productive efficiency analysis of Chinese steel firms: an application of data envelopment analysis. Ph.D: Dissertation, West Virginia University; 1999a.
Chen TY. Interpreting technical efficiency and cross-efficiency ratings in power distribution districts. Pac Asian J Energ. 1999b;9(1):31–43.
Chen TY, Yu OS. Performance evaluation of selected US utility commercial lighting demand-side management programs. J Assoc Energy Eng. 1997;94(4):50–66.
Chismar WG. Assessing the economic impact of information systems technology on organizations. Ph.D: Dissertation, Carnegie-Mellon University; 1986.
Chitkara P. A data envelopment analysis approach to evaluation of operational inefficiencies in power generating units: a case study of Indian power plants. IEEE Trans Power Syst. 1999;14 (2):419–25.
Chu X, Fielding GJ, Lamar B. Measuring transit performance using data envelopment analysis. Transport Res Part A: Pol Pract. 1992;26(3):223–30.
Clarke RL. Evaluating USAF vehicle maintenance productivity over time. Decis Sci. 1992;23 (2):376–84.

Clarke RL. 1988, Effects of Repeated Applications of Data Envelopment Analysis on Efficiency of Air Force Vehicle Maintenance Units in the Tactical Air Command and a Test for the Presence of Organizational Slack Using Rajiv Banker's Game Theory Formulations, Ph.D. Dissertation, Graduate School of Business, University of Texas.

Clarke RL, Gourdin KN. Measuring the efficiency of the logistics process. J Bus Logist. 1991;12 (2):17–33.

Co HC, Chew KS. Performance and R&D expenditures in American and Japanese manufacturing firms. Int J Prod Res. 1997;35(12):3333–48.

Collier D, Storbeck J. Monitoring of continuous improvement performance using data envelopment analysis. Proc Dec Sci Inst. 1993;3:1925–7.

Cook WD, Johnston DA. Evaluating alternative suppliers for the development of complex systems: a multiple criteria approach. J Oper Res Soc. 1991;43(11):1055–61.

Cook WD, Johnston DA, McCutcheon D. Implementation of robotics: identifying efficient implementors. Omega: Int J Manag Sci. 1992;20(2):227–39.

Cook WD, Roll Y, Kazakov A. A DEA model for measuring the relative efficiency of highway maintenance patrols. Infor. 1990;28(2):113–24.

Cook WD, Kazakov A, Roll Y. On the measuring and monitoring of relative efficiency of highway maintenance patrols. In: Charnes A, Cooper WW, Lewin A, Seiford L, editors. Data envelopment analysis: theory, methodology and applications. Norwell, MA: Kluwer; 1994.

Cook WD, Kazakov A, Roll Y, Seiford LM. A data envelopment approach to measuring efficiency: case analysis of highway maintenance patrols. J Soc Econ. 1991;20(1):83–103.

Cook WD, Kazakov A, Persaud BN. Prioritizing highway accident sites: a data envelopment analysis model. J Oper Res Soc. 2001;52(3):303–9.

Cooper WW. OR/MS: where it's been. Where it should be going? J Oper Res Soc. 1999;50:3–11.

Cooper WW, Park KSGYu. An illustrative application of IDEA (imprecise data envelopment analysis) to a Korean mobile telecommunication company. Oper Res. 2001;49(6):807–20.

Cooper WW, Seiford L, Tone K. Data envelopment analysis: a comprehensive text with models. Applications: References and DEA-Solver Software, Kluwer Academic Publishers, Boston; 2000.

Cooper WW, Sinha KK, Sullivan RS. Measuring complexity in high-technology manufacturing: indexes for evaluation. Interfaces. 1992;4(22):38–48.

Coyle RG. Systems dynamics modeling: a practical approach. 1st ed. London, Great Britain: Chapman & Hall; 1996.

Cowie J, Asenova D. Organization form, scale effects and efficiency in the British Bus industry. Transportation. 1999;26(3):231–48.

Criswell DR, Thompson RG. Data envelopment analysis of space and terrestrially based large commercial power systems for earth: a prototype analysis of their relative economic advantages. Solar Energy. 1996;56(1):119–31.

Debreu G. The coefficient of resource utilization. Econometrica. 1951;19(3):273–92.

Deprins, D., Simar, L. and H. Tulkens, 1984, Measuring Labor-Efficiency in Post Offices, in Marchand, M, Pestieau, P. and H. Tulkens, editors, The Performance of Public Enterprises: Concepts and Measurement, Elsevier Science Publishers B.V. (North Holland).

Dervaux B, Kerstens K, Vanden Eeckaut P. Radial and Non-radial static efficiency decompositions: a focus on congestion management. Transport Res-B. 1998;32(5):299–312.

Doyle JR, Green RH. Comparing products using data envelopment analysis. Omega: Int J Manag Sci. 1991;19(6):631–8.

Dubois D, Prade H. Fuzzy sets and statistical data. Eur J Oper Res. 1986;25:345–56.

Dyson RG, Allen R, Camanho AS, Podinovski VV, Sarrico CS, Shale EA. Pitfalls and protocols in DEA. Eur J Oper Res. 2001;132:245–59.

Ewing R. Measuring transportation performance. Transp Q. 1995;49(1):91–104.

Färe R, Grosskopf S. Intertemporal production frontiers: with dynamic DEA. Boston, MA: Kluwer; 1996.

Färe R, Grosskopf S, Lovell CAK. Production frontiers. Cambridge, MA: Cambridge University Press; 1994.
Färe R, Lovell CAK. Measuring the technical efficiency of production. J Econ Theor. 1978;19 (1):150–62.
Färe R, Primont D. Multi-output production and duality: theory and applications. Boston, MA: Kluwer Academic Publishers; 1995.
Färe R, Grosskopf S. Network DEA. Soc Econ PlannSci. 2000;34:35–49.
Färe R, Grosskopf S, Logan J. The comparative efficiency of western coal-fired steam electric generating plants; 1977–1979. Eng Costs Prod Econ. 1987;11:21–30.
Färe R, Grosskopf S, Pasurka C. Effects on relative efficiency in electric power generation Due to environmental controls. Resour Energ. 1986;8:167–84.
Färe R, Grosskopf S, Lovell CAK, Pasurka C. Multilateral productivity comparisons when some outputs are undesirable: a nonparametric approach. Rev Econ Stat. 1989;71(1):90–8.
Farrell MJ. The measurement of productive efficiency. J Roy Stat Soc, Series A (General). 1957;120(3):253–81.
Ferrier GD, Hirschberg JG. Climate control efficiency. Energ J. 1992;13(1):37–54.
Fielding GJ. Managing public transit strategically. San Francisco, CA: Jossey-Bass Publishers; 1987.
Fisher, 1997, An Integrated Methodology for Assessing Medical Waste Treatment Technologies (Decision Modeling), D. ENG., Southern Methodist University.
Forsund FR, Hernaes E. A comparative analysis of ferry transport in Norway. In: Charnes A, Cooper WW, Lewin A, Seiford L, editors. *Data envelopment analysis: theory, methodology and applications*. Norwell, MA: Kluwer Academic Publishers; 1994.
Forsund F, Kittelsen S. Productivity development of Norwegian electricity distribution utilities. Resource Energ Econ. 1998;20(3):207–24.
Forrester JW. Industrial dynamics. Cambridge, MA: MIT Press; 1961.
Forrester JW. Principles of systems. Cambridge, MA: MIT; 1968.
Fried H, Lovell CAK, Schmidt S, editors. The measurement of productive efficiency. Oxford: Oxford University Press; 1993.
Frisch, R. On the Notion of Equilibrium and Disequilibrium. Rev Econ Stud. 1935–1936; 100–106.
Gathon H-J. Indicators of partial productivity and technical efficiency in European transit sector. Annals of Public and Co-operative Econ. 1989;60(1):43–59.
Gillen D, Lall A. Developing measures of airport productivity and performance: an application of data envelopment analysis. Transport Res Part E-Logist Transport Rev. 1997;33(4):261–73.
Giokas DI, Pentzaropoulos GC. Evaluating the relative efficiency of large-scale computer networks-an approach via data envelopment analysis. Appl Math Model. 1995;19(6):363–70.
Girod O. Measuring technical efficiency in a fuzzy environment. Ph.D: Dissertation, Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University; 1996.
Girod O, Triantis K. The evaluation of productive efficiency using a fuzzy mathematical programming approach: the case of the newspaper preprint insertion process. IEEE Trans Eng Manag. 1999;46(4):1–15.
Golany B, Roll Y, Rybak D. Measuring efficiency of power-plants in Israel by data envelopment analysis. IEEE Trans Eng Manag. 1994;41(3):291–301.
Golany B, Roll Y. Incorporating standards via data envelopment analysis. In: Charnes A, Cooper WW, Lewin A, Seiford L, editors. Data envelopment analysis: theory, methodology and applications. Norwell, MA: Kluwer; 1994.
Haas DA. Evaluating the efficiency of municipal reverse logistics channels: an application of data envelopment analysis (solid waste disposal). Ph.D. Dissertation: Temple University; 1998.
Hayes KE, Ratick S, Bowen WM, CummingsSaxton J. Environmental decision models: US experience and a New approach to pollution management. Environ Int. 1993;19:261–75.

Hjalmarsson L, Odeck J. Efficiency of trucks in road construction and maintenance: an evaluation with data envelopment analysis. Comput Oper Res. 1996;23(4):393–404.
Hollingsworth KB. A warehouse benchmarking model utilizing frontier production functions (data envelopment analysis). Dissertation, Georgia Institute of Technology: Ph.D; 1995.
Hoopes B, Triantis K. Efficiency performance, control charts and process improvement: complementary measurement and evaluation. IEEE Trans Eng Manag. 2001;48(2):239–53.
Hoopes B, Triantis K, Partangel N. The relationship between process and manufacturing plant performance: a goal programming approach. Int J Oper Quant Manag. 2000;6(4):287–310.
Hougaard J. Fuzzy scores of technical efficiency. Eur J Oper Res. 1999;115:529–41.
Husain N, Abdullah M, Kuman S. Evaluating public sector efficiency with data envelopment analysis (DEA): a case study in road transport department, Selangor, Malaysia. Total Qual Manag. 2000;11(4/5):S830–6.
IEEE Transactions on Fuzzy Systems (1994), February 1994, volume 2, number 1, pp. 16–45.
Inuiguchi M, Tanino T. Data envelopment analysis with fuzzy input and output data. Lect Notes Econ Math Syst. 2000;487:296–307.
Johnson, A. L. and L. F. McGinnis, 2010, Productivity Measurement in the Warehousing Industry, forthcoming, IEE Transactions.
Kabnurkar, A., 2001, Math Modeling for Data Envelopment Analysis with Fuzzy Restrictions on Weights, M.S. Thesis, *Virginia Tech, Department of Industrial and Systems Engineering*, Falls Church, VA.
Kao C, Liu ST. Fuzzy efficiency measures in data envelopment analysis. Fuzzy Set Syst. 2000;113 (3):427–37.
Karsak, E.E., 1999, DEA-based Robot Selection Procedure Incorporating Fuzzy Criteria Values, *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 1, I-1073-I-1078, IEEE.
Kazakov A, Cook WD, Roll Y. Measurement of highway maintenance patrol efficiency: model and factors. Transp Res Rec. 1989;1216:39–45.
Kemerer CF. 1987, Measurement of Software development, Ph.D. Dissertation, Graduate School of Industrial Administration, Carnegie-Mellon, University.
Kemerer, C.F., 1988, Production Process Modeling of Software Maintenance Productivity, *Proceedings of the IEEE Conference on Software Maintenance*, p. 282, IEEE Computer Society Press, Washington, DC, USA.
Kerstens K. Technical efficiency measurement and explanation of French urban transit companies. Transport Res-A. 1996;30(6):431–52.
Khouja M. The Use of data envelopment analysis for technology selection. Comput Ind Eng. 1995;28(1):123–32.
Kim SH, Park C-G, Park K-S. Application of data envelopment analysis in telephone offices evaluation with partial data. Comput Oper Res. 1999;26(1):59–72.
Kleinsorge IK, Schary PB, Tanner RD. Evaluating logistics decisions. Int J Phys Distrib Mater Manag. 1989;19(12):3–14.
Kopp RJ. The measurement of productive efficiency: a reconsideration. Q J Econ. 1981;96:476–503.
Koopmans T. Analysis of production as an efficient combination of activities, Activity analysis of production and allocation, New Haven, Yale University Press, 1951. p. 3–97.
Kumar M. 2002, A Preliminary Examination of the use of DEA (Data Envelopment Analysis) for Measuring Production Efficiency of a Set of Independent Four Way Signalized Intersections in a Region, MS Thesis, Virginia Polytechnic Institute and State University, Department of Civil Engineering, Advanced Transportation Systems.
Kumar C, Sinha BK. Efficiency based decision rules for production planning and control. Int J Syst Sci. 1998;29(11):1265–80.
Land KC, Lovell CAK, Thore S. Chance-constrained efficiency analysis. Manag Decis Econ. 1993;14:541–53.

Laviolette M, Seaman JW, Barrett JD, Woodall WH. A Probabilistic and Statistical View of Fuzzy Methods. Technometrics. 1995;37:249–61.
Laviolette M, Seaman JW. Unity and diversity of fuzziness-from a probability viewpoint. IEEE Trans Fuzzy Syst. 1994;2(1):38–42.
Lebel LG. 1996, Performance and efficiency evaluation of logging contractors using data envelopment analysis, Ph.D. Dissertation, Virginia Polytechnic Institute and State University.
Lee SK, Migi G, Kim JW Multi-criteria decision making for measuring relative efficiency of greenhouse gas technologies, AHP/DEA hybrid model approach, Eng Lett. 2008;16:4, EL_4_05.
Lelas V. 1998, Chance constrained models for air pollution monitoring and control (Risk Management), Ph.D. Dissertation, The University of Texas at Austin.
Li Z, Liao H, Coit D. A Two stage approach for multi-objective decision making with applications to system reliability engineering. Reliab Eng Saf. 2009;94:1585–92.
Liangrokapart J. 2001, Measuring and enhancing the performance of closely linked decision making units in supply chains using customer satisfaction data, Ph. Dissertation, Clemson University.
Linton JD, Cook WD. Technology implementation: a comparative study of Canadian and US factories. Infor. 1998;36(3):142–50.
Liu S-T. A fuzzy DEA/AR approach to the selection of flexible manufacturing systems. Comput Ind Eng. 2008;54:66–76.
Löthgren M, Tambour M. Productivity and customer satisfaction in Swedish pharmacies: a DEA network model. Eur J Oper Res. 1999;115:449–58.
Lovell CAK (1997), "What a Long Strange Trip It's Been," Fifth European Workshop on Efficiency and Productivity Analysis, October 9–11, 1997, Copenhagen, Denmark.
Mahmood MA, Pettingell KJ, Shaskevich AI. Measuring productivity of software projects-a data envelopment analysis approach. Decis Sci. 1996;27(1):57–80.
Martinez M. 2001, Transit Productivity Analysis in Heterogeneous Conditions Using Data Envelopment Analysis with an Application to Rail Transit, Ph.D. Dissertation, New Jersey Institute of Technology.
Majumdar SK. Does technology adoption Pay-electronic switching patterns and firm-level performance in US telecommunications. Research Policy. 1995;24(5):803–22.
Majumdar SK. Incentive regulation and productive efficiency in the US telecommunication industry. J Bus. 1997;70(4):547–76.
McMullen PR, Frazier GV. Using simulation and data envelopment analysis to compare assembly line balancing solutions. J Product Anal. 1999;11(2):149–68.
McMullen, P.R. and G.V. Frazier, 1996, Assembly Line Balancing Using Simulation and Data Envelopment Analysis, *Proceedings of the Annual Meeting-Decision Sciences Institute*, volume 3.
Meeusen W, Vanden Broeck J. Efficiency estimation from Cobb-Douglas production functions with composed error. Int Econ Rev. 1977;18(2):435–44.
Miyashita T, Yamakawa H. A study on the collaborative design using supervisor system. JSME Int J Series C. 2002;45(1):333–41.
Morita H, Kawasakim T, Fujii S. 1996, Two-objective set division problem and its application to production cell assignment
Nakanishi YJ, Norsworthy JR. Assessing Efficiency of Transit Service. IEEE International Engineering Management Conference, IEEE, Piscataway, NJ, USA; 2000; p. 133–140.
Nijkamp P, Rietveld P. Multi-objective multi-level policy models: an application to regional and environmental planning. Eur Econ Rev. 1981;15:63–89.
Nolan JF. Determinants of productive efficiency in urban transit. Logist Transport Rev. 1996;32 (3):319–42.
Nozick LK, Borderas H, Meyburg AH. Evaluation of travel demand measures and programs: a data envelopment analysis approach. Transport Res-A. 1998;32(5):331–43.

Obeng K, Benjamin J, Addus A. Initial analysis of total factor productivity for public transit. Transp Res Rec. 1986;1078:48–55.
O'Connor, J. and I. McDermott, 1997, The art of systems thinking: essential skills for creativity and problem solving, Thorsons
Odeck J. 1993, Measuring Productivity Growth and Efficiency with Data Envelopment Analysis: An Application on the Norwegian Road Sector, Ph.D. Dissertation, Department of Economics, University of Goteborg, Goteborg, Sweden.
Odeck J. Evaluating efficiency of rock blasting using data envelopment analysis. J Transport Eng-ASCE. 1996;122(1):41–9.
Odeck J. Assessing the relative efficiency and productivity growth of vehicle inspection services: an application of DEA and malmquist indices. Eur J Oper Res. 2000;126(3):501–14.
Odeck J, Hjalmarsson L. The performance of trucks-an evaluation using data envelopment analysis. Transp Plan Technol. 1996;20(1):49–66.
Olesen OB, Petersen NC. Chance constrained efficiency evaluation. Manag Sci. 1995;41 (3):442–57.
Otis PT. 1999, Dominance Based Measurement of Environmental Performance and Productive Efficiency of Manufacturing, Ph.D. Dissertation, Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University.
Ozbek M. 2007, Development of Comprehensive Framework for the Efficiency Measurement of Road Maintenance Strategies Using Data Envelopment Analysis, Ph.D. in Civil Engineering. Virginia Polytechnic Institute and State University.
Ozbek M, de la Garza J, Triantis K. Data and modeling issues faced during the efficiency measurement of road maintenance using data envelopment analysis. J Infra Syst, 2010a; ASCE, 16(1), 21–30, (supported by NSF grant # 0726789).
Ozbek M, de la Garza J, Triantis K. Efficiency measurement of bridge maintenance using data envelopment analysis, J Infra Syst 2010b;16(1), 31–39 (supported by NSF grant # 0726789).
Papahristodoulou C. A DEA model to evaluate car efficiency. Appl Econ. 1997;29 (11):14913–1508.
Paradi JC, Reese DN, Rosen D. Applications of DEA to measure of software production at two large Canadian banks. Ann Oper Res. 1997;73:91–115.
Paradi JC, Smith S, Schaffnit-Chatterjee C. Knowledge worker performance analysis using DEA: an application to engineering design teams at bell Canada. IEEE Trans Eng Manag. 2002;49 (2):161–72.
Peck MW, Scheraga CA, Boisjoly RP. Assessing the relative efficiency of aircraft maintenance technologies: an application of data envelopment analysis. Transport Res Part A-Policy Pract. 1998;32(4):261–9.
Peck MW, Scheraga CA, Boisjoly RP. 1996, The utilization of data envelopment analysis in benchmarking aircraft maintenance technologies, Proceedings of the 38th Annual Meeting-Transportation Research Forum, 1, 294–303.
Pedraja-Chaparro FSalinas-Jiménez, Smith P. On the quality of the data envelopment analysis. J Oper Res Soc. 1999;50:636–44.
Polus A, Tomecki AB. Level-of-service framework for evaluating transportation system management alternatives. Transp Res Rec. 1986;1081:47–53.
Pratt RH, Lomax TJ. Performance measures for multi modal transportation systems. Transp Res Rec. 1996;1518:85–93.
Ramanathan R. Combining indicators of energy consumption and CO2 emissions: a cross-country comparison. Int J Global Energ Issues. 2002;17(3):214–27.
Ray SC, Hu XW. On the technically efficient organization of an industry: a study of US airlines. J Product Anal. 1997;8(1):5–18.
Resti A. Efficiency measurement for multi-product industries: a comparison of classic and recent techniques based on simulated data. Eur J Oper Res. 2000;121(3):559–78.
Richmond B. The "thinking" in systems thinking: seven essential skills. Waltham, MA: Pegasus Communications, Inc.; 2000.

Ross A, Venkataramanan MA. 1998, Multi Commodity-Multi Echelon Distribution Planning: A DSS Approach with Application, Proceedings of the Annual Meeting-Decision Sciences.
Rouse P, Putterill M, Ryan D. Towards a general managerial framework for performance measurement: a comprehensive highway maintenance application. J Product Anal. 1997;8 (2):127–49.
Rouse P, Chiu T. Towards optimal life cycle management in a road maintenance setting using DEA. Eur J Oper Res. 2008. doi:10.1016/j.ejor.2008.02.041.
Ryus P, Ausman J, Teaf D, Cooper M, Knoblauch M. Development of Florida's transit level-of-service indicator. Transp Res Rec. 2000;1731:123–9.
Samuelson PA. Foundations of economic analysis. Cambridge, MA: Harvard University Press; 1947.
Sarkis J. An empirical analysis of productivity and complexity for flexible manufacturing systems. Int J Prod Econ. 1997a;48(1):39–48.
Sarkis J. Evaluating flexible manufacturing systems alternatives using data envelopment analysis. Eng Econ. 1997b;43(1):25–47.
Sarkis J. Methodological framework for evaluating environmentally conscious manufacturing programs. Comput Ind Eng. 1999;36(4):793–810.
Sarkis J, Cordeiro J. 1998. Empirical evaluation of environmental efficiencies and firm performance: pollution prevention versus end-of-pipe practice, *Proceedings of the Annual Meeting-Decision Sciences Institute*.
Sarkis J, Talluri S. Efficiency evaluation and business process improvement through internal benchmarking. Eng Evalu Cost Anal. 1996;1:43–54.
Sarkis J, Talluri S. A decision model for evaluation of flexible manufacturing systems in the presence of both cardinal and ordinal factors. Int J Prod Res. 1999;37(13):2927–38.
Sarkis J, Weinrach J. Using data envelopment analysis to evaluate environmentally conscious waste treatment technology. J Clean Prod. 2001;9(5):417–27.
Scheraga CA, Poli PM. 1998, Assessing the Relative Efficiency and Quality of Motor Carrier Maintenance Strategies: An Application of Data Envelopment Analysis, Proceedings of the 40th Annual Meeting-Transportation Research Forum, Transportation Research Forum, 1, 163–185.
Seaver B, Triantis K. The implications of using messy data to estimate production frontier based technical efficiency measures. J Bus Econ Stat. 1989;7(1):51–9.
Seaver B, Triantis K. The impact of outliers and leverage points for technical efficiency measurement using high breakdown procedures. Manag Sci. 1995;41(6):937–56.
Seaver B, Triantis K. A fuzzy clustering approach used in evaluating technical efficiency measures in manufacturing. J Product Anal. 1992;3:337–63.
Seaver B, Triantis K, Reeves C. Fuzzy selection of influential subsets in regression. Technometrics. 1999;41(4):340–51.
Seaver B, Triantis K, Hoopes B. Efficiency performance and dominance in influential subsets: an evaluation using fuzzy clustering and pair-wise dominance. J Product Anal. 2004;21:201–20.
Seiford L. 1999, An Introduction to DEA and a Review of Applications in Engineering, NSF Workshop on Engineering Applications of DEA, Union College, NY, December, 1999.
Sengupta JK. Data envelopment analysis for efficiency measurement in the stochastic case. Comput Oper Res. 1987;14(2):117–29.
Sengupta JK. A fuzzy systems approach in data envelopment analysis. Comput Math Appl. 1992;24(8/9):259–66.
Shafer SM, Bradford JW. Efficiency measurement of alternative machine components grouping solutions via data envelopment analysis. IEEE Trans Eng Manag. 1995;42(2):159–65.
Shao B. 2000, Investigating the Value of Information Technology in Productive Efficiency: An Analytic and Empirical Study, Ph.D. Dissertation, State University of New York in Buffalo.
Shash, A.A.H., 1988, A Probabilistic Model for U.S. Nuclear Power Construction Times, Ph.D. Dissertation, Department of Civil Engineering, University of Texas.
Shephard RW. Cost and production functions. Princeton, NJ: Princeton University Press; 1953.

Shephard RW. Theory of cost and production functions. Princeton, NJ: Princeton University Press; 1970.
Sheth, N., 1999, Measuring and Evaluating Efficiency and Effectiveness Using Goal Programming and Data Envelopment Analysis in a Fuzzy Environment, M.S. Thesis, *Virginia Tech, Department of Industrial and Systems Engineering*, Falls Church, VA.
Sheth C, Triantis K, Teodorović D. The measurement and evaluation of performance of urban transit systems: the provision of Bus service along different routes. Transport Res: Part E. 2007;43:453–78.
Shewhart WA. 1980, Economic control of quality in manufacturing, D. Van Nostrand, New York (Republished by the American Society for Quality Control, Milwaukee, WI, 1980).
Sinha KK, 1991, Models for evaluation of complex technological systems: strategic applications in high technology manufacturing, Ph.D. Dissertation, Graduate School of Business, University of Texas.
Sjvgren S. 1996, Efficient Combined Transport Terminals-A DEA Approach, Department of Business Administration, University of Götenberg.
Soloveitchik D, Ben-Aderet N, Grinman M, Lotov A. Multiobjective optimization and marginal pollution abatement cost in the electricity sector – an Israeli case study. Eur J Oper Res. 2002;140(3):571–83.
Smith JK. 1996, The Measurement of the Environmental Performance of Industrial Processes: A Framework for the Incorporation of Environmental Considerations into Process Selection and Design, Ph.D. Dissertation, Duke University.
Sterman JD. Business dynamics: systems thinking and modeling for a complex world. Boston, MA: Irwin McGraw-Hill; 2000.
Storto C.L., 1997, Technological Benchmarking of Products Using Data Envelopment Analysis: An Application to Segments A' and B' of the Italian Car Market, *Portland International Conference on Management of Engineering and Technology*, D.F. Kocaoglu and T.R. Anderson, editors, 783–788.
Sueyoshi T. Tariff structure of Japanese electric power companies: an empirical analysis using DEA. Eur J Oper Res. 1999;118(2):350–74.
Sueyoshi T, Machida H, Sugiyama M, Arai T, Yamada Y. Privatization of Japan national railways: DEA time series approaches. J Oper Res Soc Jpn. 1997;40(2):186–205.
Sun S. Assessing computer numerical control machines using data envelopment analysis. Int J Prod Res. 2002;40(9):2011–39.
Talluri S, Baker RC, Sarkis J. A framework for designing efficient value chain networks. Int J Prod Econ. 1999;62(1–2):133–44.
Talluri S, Huq F, Pinney WE. Application of data envelopment analysis for cell performance evaluation and process improvement in cellular manufacturing. Int J Prod Res. 1997;35 (8):2157–70.
Talluri S, Sarkis J. Extensions in efficiency measurement of alternate machine component grouping solutions via data envelopment analysis. IEEE Trans Eng Manag. 1997;44 (3):299–304.
Talluri S, Yoon KP. A cone-ratio DEA approach for AMT justification. Int J Prod Econ. 2000;66:119–29.
Talluri S. 1996, A Methodology for Designing Effective Value Chains: An Integration of Efficient Supplier, Design, Manufacturing, and Distribution Processes (Benchmarks), Ph.D. Dissertation, The University of Texas at Arlington.
Talluri, S., 1996, Use of Cone-Ratio DEA for Manufacturing Technology Selection, Proceedings of the Annual Meeting-Decision Sciences Institute.
Technometrics, 1995, volume 37, no. 3, August 1995, pp. 249–292.
Teodorović D. Invited review: fuzzy sets theory applications in traffic and transportation. Eur J Oper Res. 1994;74:379–90.
Teodorović D. Fuzzy logic systems for transportation engineering: the state of the Art. Transp Res. 1999;33A:337–64.

Teodorović D, Vukadinovic K. Traffic control and transport planning: a fuzzy sets and neural networks approach. Boston, MA: Kluwer; 1998.
Thanassoulis E, Dyson RG. Estimating preferred target input-output levels using data envelopment analysis. Eur J Oper Res. 1992;56:80–97.
Thompson RG, Singleton Jr FD, Thrall RM, Smith BA. Comparative site evaluation for locating a high-energy physics Lab in Texas. Interfaces. 1986;16(6):35–49.
Tone K, Sawada T. 1991, An Efficiency Analysis of Public Vs. Private Bus Transportation Enterprises, Twelfth IFORS International Conference on Operational Research, 357–365.
Tofallis C. Input efficiency profiling: an application to airlines. Comput Oper Res. 1997;24 (3):253–8.
Tran A, Womer K. Data envelopment analysis and system selection. Telecomm Rev 1993; 107–115.
Triantis K. 1984, Measurement of Efficiency of Production: The Case of Pulp and Linerboard Manufacturing, Ph.D. Dissertation, Columbia University.
Triantis K. 1987, Total and partial productivity measurement at the plant level: empirical evidence for linerboard manufacturing, productivity management frontiers – I, edited by D. Sumanth, Elsevier Science Publishers, Amsterdam, 113–123.
Triantis K. 1990, An assessment of technical efficiency measures for manufacturing plants, People and product management in manufacturing, advances in industrial engineering, No. 9, edited by J. A. Edosomwan, Elsevier Science Publishers, Amsterdam, 149–166.
Triantis K. 2003, Fuzzy Non-Radial DEA Measures of Technical Efficiency, forthcoming, International Journal of Automotive Technology and Management.
Triantis K, Girod O. A mathematical programming approach for measuring technical efficiency in a fuzzy environment. J Product Anal. 1998;10:85–102.
Triantis K, Otis P. A dominance based definition of productive efficiency for manufacturing taking into account pollution prevention and recycling, forthcoming. Eur J Oper Res. 2003.
Triantis K, Medina-Borja A. 1996, Performance Measurement: The Development of Outcome Objectives: Armed Forces Emergency Services," *American Red Cross*, Chapter Management Workbook, Armed Forces Emergency Services, System Performance Laboratory.
Triantis K, NcNelis R. 1995, The Measurement and Empirical Evaluation of Quality and Productivity for a Manufacturing Process: A Data Envelopment Analysis (DEA) Approach, *Flexible Automation and Intelligent Manufacturing-5th International Conference*, Schraft, R.D., editor, 1134–1146, Begell House Publishers.
Triantis K, Vanden Eeckaut P. Fuzzy pairwise dominance and implications for technical efficiency performance assessment. J Product Anal. 2000;13(3):203–26.
Triantis, K., Coleman, G., Kibler, G., and Sheth, N., 1998, Productivity Measurement and Evaluation in the United States Postal Service at the Processing and Distribution Center Level, System Performance Laboratory, distributed to the *United States Postal Service*.
Triantis K, Sarangi S, Kuchta D. Fuzzy pair-wise dominance and fuzzy indices: an evaluation of productive performance. Eur J Oper Res. 2003;144:412–28.
Tyteca D. 1995, Linear Programming Models for the Measurement of Environmental Performance of Firms – Concepts and Empirical Results, Intitut d'Administration et de Gestion Université Catholique de Louvain, Place des Doyens, 1, B-1348, Louvain-la-Neuve, Belgium, September.
Tyteca D. On the measurement of the environmental performance of firms – a literature review and a productive efficiency perspective. J Environ Manag. 1996;46:281–308.
Uri ND. Changing productive efficiency in telecommunications in the United States. Int J Prod Econ. 2001;72(2):121–37.
Vaneman W. Evaluating performance in a complex and dynamic environment. Ph.D: Dissertation, Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University; 2002.
Vaneman W, Triantis K. The dynamic production axioms and system dynamics behaviors: the foundation for future integration. J Product Anal. 2003;19(1):93–113.

Vaneman W, Triantis K. Evaluating the productive efficiency of dynamical systems. IEEE Trans Eng Manag. 2007;54(3):600–12.
Vargas VA, Metters R. Adapting Lot-sizing techniques to stochastic demand through production scheduling policy. IIE Trans. 1996;28(2):141–8.
Wang B, Zhang Q, Wang F. Using DEA to evaluate firm productive efficiency with environmental performance. Control Dec. 2002;17(1):24–8.
Wang CH, Gopal RD, Zionts S. Use of data envelopment analysis in assessing information technology impact on firm performance. Ann Oper Res. 1997;73:191–213.
Wang CH, 1993, The Impact of Manufacturing Performance on Firm Performance, the Determinants of Manufacturing Performance and the Shift of the Manufacturing Efficiency Frontier, Ph.D. Dissertation, State University of New York in Buffalo.
Ward P, Storbeck JE, Magnum SL, Byrnes PE. An analysis of staffing efficiency in US manufacturing: 1983 and 1989. Ann Oper Res. 1997;73:67–90.
Wilson PW. Detecting outliers in deterministic nonparametric frontier models with multiple outputs. J Bus Econ Stat. 1993;11:319–23.
Wilson PW. Detecting influential observations in data envelopment analysis. J Product Anal. 1995;6:27–45.
Wolstenholme EF. System enquiry: a system dynamics approach. New York, NY: John Wiley & Sons; 1990.
Wu, T. Fowler, J., Callarman, T., and A. Moorehead, 2006, Multi-Stage DEA as a Measurement of Progress in Environmentally Benign Manufacturing, *Flexible Automation and Intelligent Manufacturing, FAIM 2006*, Limerick Ireland.
Wu L, Xiao C. Comparative sampling research on operations management in machine tools industry between china and the countries in Western Europe (in Chinese). J Shanghai Inst Mech Eng. 1989;11(1):61–7.
Ylvinger S. Industry performance and structural efficiency measures: solutions to problems in firm models. Eur J Oper Res. 2000;121(1):164–74.
Zadeh LA. Fuzzy sets. Inf Control. 1965;8:338–53.
Zeng G. Evaluating the efficiency of vehicle manufacturing with different products. Ann Oper Res. 1996;66:299–310.
Zhu J, Chen Y. Assessing textile factory performance. J Syst Sci Syst Eng. 1993;2(2):119–33.

Chapter 15
Applications of Data Envelopment Analysis in the Service Sector

Necmi K. Avkiran

Abstract The service sector holds substantial challenges for productivity analysis because most service delivery is often heterogeneous, simultaneous, intangible, and perishable. Nevertheless, the prospects for future studies are promising as we gently push the data envelopment analysis research envelope by using more innovative research designs that may include synergistic partnerships with other methods and disciplines, as well as delve deeper into the sub-DMU network of organizations. This chapter is dedicated to providing a selection of applications in the service sector with a focus on building a conceptual framework, research design, and interpreting results. Given the expanding share of the service sector in gross domestic products of many countries, the twenty-first century will continue to provide fertile grounds for research in the service sector.

Keywords Universities • Hotels • Real estate agents • Commercial banks • Technical and scale inefficiencies • Profitability versus inefficiency • Network DEA

15.1 Introduction

In a life span exceeding 30 years, researchers have produced many creative applications of data envelopment analysis (DEA) that span practically all sectors and industries therein. For example, it is possible to find applications of DEA from fish farms in China (Sharma et al. 1999) to health care management (see the recent publications Amado and Dyson 2009; Vitikainen et al. 2009). Nevertheless,

N.K. Avkiran (✉)
UQ Business School, The University of Queensland, Brisbane, QLD 4072, Australia
e-mail: n.avkiran@business.uq.edu.au

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_15,

the service sector presents substantial challenges in productivity analysis because of the nature of most service delivery which can be described as heterogeneous, simultaneous, intangible, and perishable. A survey of articles (excluding conference proceedings) in the Web of Science data base reveals that DEA has been applied in a variety of service sector industries, including banking and insurance (175 articles), health care (75 articles), universities (68 articles), tourism (11 articles), and real estate (3 articles). A larger selection of subject areas available on the Web of Science most likely to include the services sector indicate 605 articles, out of a total of 2,618 articles on DEA (as at February 11, 2010). Subject areas related to the service sector include anesthesiology, communication, computer science, education, health care sciences, health policy, hospitality, information science, law, medical informatics, occupational health, planning and development, psychiatry, psychology, public administration, social sciences, tourism, and so on.

Popularity of DEA remains strong as judged by the multiple academic symposia dedicated to the discipline every year and the DEA streams we encounter in annual productivity workshops. This is partly due to the intuitive and forgiving nature of this nonparametric technique which presents a short learning curve for its users, as well as the compelling nature of efficiency analysis which is part of modern business and public sector organizations operating in a globally competitive environment. Directions for future studies are many fold, including opening the black-box of production (illustrated at the end of this chapter), dynamic DEA, rationalizing inefficiency, the productivity dilemma, and synergistic partnerships with other methods such as goal programming, all of which promise a new wave of applications across a wide range of disciplines. The rest of this chapter is dedicated to providing a selection of DEA's applications in the service sector covering universities, hotels, real estate agents, and banks, with substantial space dedicated to the process of selecting appropriate inputs and outputs. Readers are also referred to Avkiran (2006) for additional examples on productivity analysis in the service sector using DEA. Highlights of the conclusions from various cases discussed in the remainder of this chapter are presented next.

The application of DEA to universities suggests that decomposing technical efficiency scores into pure technical efficiency, and scale efficiency can provide guidance on what can be achieved in the short versus long term. For example, if most of the observed inefficiency is due to a small scale, then that university can consider expansion. While this exercise can be fast-tracked through in-market mergers in the case of profit-making organizations, it is a much more bureaucratic and lengthy process for universities which normally have to extensively consult key stakeholders such as Federal and State governments. On the other hand, pure technical inefficiency can often be addressed in a shorter time period. There are also potential problems in collecting data on research income and research outcome in the context of universities. The long and variable lead times between publications and payment of research income means universities cannot easily collect timely and relevant data for a particular year. While the performance

of universities is increasingly evaluated on quantifiable variables, higher education still retains certain characteristics that set it apart from other types of organizations. This continues to complicate selection of inputs/outputs and comparisons across institutions that satisfy all stakeholders.

The application of DEA in the hospitality industry discussed in this chapter suggests that some of the hotels could significantly reduce their bed numbers and part-time staff numbers while simultaneously increasing outputs of revenue and room rates. Yet, particularly in high-class hotels where there is fierce competition based on service quality and guests expect personal service on demand, it may not be possible to significantly reduce the number of staff without lowering service quality. DEA enables management to integrate dissimilar multiple inputs and outputs, for example, accounting measures and customer service quality, to make comparisons that would otherwise not be possible. On an everyday basis, management can use DEA to test their established knowledge of existing operations and investigate further when conflicting findings emerge. More often than not, the process of DEA encourages management to identify the connections between business goals and key business drivers.

The application of DEA to real estate agents indicates that some agents may appear as a profitable operation but with suboptimal efficiency. This could be due to favorable environmental conditions such as low competition in business catchment area. Other agents with efficient operations but with suboptimal profitability probably have to deal with tougher competition or a local market with a low turnover that hampers generating a higher surplus. The results further suggest that the accounting measure operating surplus and relative efficiency score correlate at the two extremes. There are cases that imply that efficient agents are not necessarily profitable operations and vice versa. While there may well be environmental factors not captured in the model that are causing the discrepancies, the lack of a strong correlation between profitability and efficiency is well documented in DEA literature. For example, according to Sherman and Zhu (2006, p. 302), "...studies of benchmarking practices with DEA have identified numerous sources of inefficiency in some of the most profitable firms...."

In the last case presented in this chapter, the application of DEA to commercial banking is motivated by the key limitation of traditional DEA, namely, its inability to reveal specific underlying sources of organizational inefficiency at the subunit level. In a global economy where most businesses compete with firms in other countries, managers keen to sustain their competitive advantage are increasingly looking for methods of efficiency analysis that can provide a more detailed insight. The case with commercial banks demonstrates network DEA – an advanced approach to DEA that has gained renewed interest in recent years. Network DEA provides access to underlying diagnostic information in a firm's divisions that would otherwise remain unknown. Applying insightful performance methodologies such as network DEA could also help in decision making involving mergers where managers typically search for synergies based on strengths of merging organizations' subunits.

15.2 Universities[1]

15.2.1 Introduction

The purpose of this section is to demonstrate the use of DEA in examining the relative efficiency of universities. The case study investigates data on the population of Australian universities (36 in all, excluding the only private institution, Bond University). Data have been collected from various government publications. Three productivity models are tested assuming output maximization and variable returns to scale.

The trend toward a user-pays system makes this study timely and relevant to the needs of decision makers in the public sector. Efficiency analysis, in one form or another, will become more widespread among universities who are increasingly made to account for public funds granted to them. As a result, the educational administrators will continue to feel more pressure to better utilize scarce resources entrusted to them. In fact, staff productivity has been central to salary negotiations for some time. Failure to make efficiency analysis a standard practice is likely to lead to less than optimal allocation of resources. It should be noted at the outset that the study does not aim to rank order the universities in the sample or create league tables.

Public sector performance indicators have been criticized for being inadequate and not conducive to analyzing efficiency (Birch and Maynard 1986; Barrow and Wagstaff 1989). Criticism leveled at public sector performance indicators list such issues as (a) focusing on inputs at the detriment of outputs, (b) an ad hoc selection of indicators, and (c) inability to distinguish inefficiency from environmental factors. Nevertheless, in recent times, DEA literature has provided solutions to how to account for environmental factors (e.g., see Avkiran 2009b; Avkiran and Thoraneenitiyan 2010). A wider discussion of the difficulties of performance measurement in public sector services can be read in Flynn (1986).

DEA is particularly suited for the job where the analyst is trying to examine the efficiency of converting multiple inputs into multiple outputs. The three models of university efficiency developed in this study are overall performance, performance on delivery of educational services, and performance on fee-paying enrolments. The findings show that Australian universities are, in general, technically and scale efficient. The largest potential improvements for Australian universities were in fee-paying enrolments as revealed by the third model. Small input slacks were found across the first two models. Majority of Australian universities had apparently been operating under a regime of decreasing returns to scale (DRS) while about one-third have been at a most productive scale size (MPSS), with a small number at increasing returns to scale (IRS).

[1] ADAPTED FROM AVKIRAN 2001.

MPSS (Banker and Thrall 1992; Banker 1984) implies that a decision-making unit (DMU)'s production of outputs is maximized per unit of inputs. That is, scale elasticity equals 1.

15.2.2 Conceptual Framework

How can the educational administrator use DEA? While DEA generates efficiency scores helping to distinguish efficient from inefficient units, a list of units sorted on efficiency scores cannot always be considered as truly rank ordered (Sherman 1988). DEA identifies a unit as either efficient or inefficient compared to other units in its reference set, where the reference set is comprised of efficient units most similar to that unit in their configuration of inputs and outputs. Knowing which efficient universities are most comparable to the inefficient university thus enables the educational administrator to better understand the relevant inefficiencies and subsequently reallocate scarce resources to improve productivity.

An overview of difficulties with some of the performance indicators in higher education could help understand the potential role of DEA in efficiency analysis. For example, Cave et al. (1991) describe a performance indicator such as *degree results* as a quality-adjusted measure of output. This indicator is typical of the ambiguity found in education performance indicators in that high achievement degree results may, for example, be due to high entry qualifications rather than effectiveness of teaching. This value added productivity indicator, described as an input- and quality-adjusted output measure, relies on differences in qualifications that cannot be valued in monetary terms.

Typical performance indicators on research include number of publications, number of citations of publications, journal impact factors, and reputational ranking. Number of publications, often interpreted as a measure of research activity, suffers from the problem of different practices across disciplines. For example, publishing in medicine may consume more time and resources and lead to fewer publications than publishing in business. There is also the difficulty of ownership since a research project started in one university will often be carried to another university when the researcher relocates. The number of citations attracted by an article has its drawbacks as well; for example, articles in mathematics have a longer shelf life than those in pharmacy (Johnes 1988).

Impact factors are not particularly problem free either. For instance, academics that devote most of their time to writing articles that are published in low-impact journals will be disadvantaged in this count regardless of the contribution of research to the body of knowledge. On the other hand, reputational ranking is bound to have considerable subjectivity and reflect a historical performance (Cave et al. 1991). To counter such concerns, as of 2010, the Australian Federal Government has taken the lead by publishing the so-called Excellence in Research for Australia or ERA ranked journal list (see http://www.arc.gov.au/era/era_journal_list.htm).

The main difficulty in selecting performance indicators in higher education is adequately capturing the interplay among the various inputs and outputs. This is where DEA offers a mathematical solution but the selection of inputs and outputs remains a critical step in efficiency analysis.

As has already been indicated, the current study develops performance models from the perspective of the educational administrator or manager. A book by Thomas (1974) discusses three ways of conceptualizing the production process (inputs and outputs) of an educational institution. These are the production functions of the psychologist, the economist, and the administrator. Briefly, the psychologist's perspective is concerned with the behavioral changes in students, including values and interpersonal skills. However, the psychologist's perspective generates less agreement on the choice of inputs and outputs compared to the other perspectives. It often utilizes subjective inputs/outputs, which are difficult to quantify. The economist's perspective concerns itself with the additional earnings generated by expenditure on schooling. Put another way, education is regarded as contributing to the economy through individuals with competencies (Lindsay 1982).

The administrator's perspective focuses on the institutional manager's attempts at allocating resources to provide services. The key outputs under this approach emanate from teaching and research activities, with research as the more difficult of the two to measure. The administrator would be primarily interested in controllable inputs. For example, in the case of staff inputs, it is appropriate to use salary as a proxy since it represents the monetary value of key inputs into the teaching and research processes (Lindsay 1982). Alternatively, a measurement can be taken with full-time equivalence (FTE).

In education, it is difficult to use market mechanisms such as profits to determine the performance of a DMU (Anderson and Walberg 1997). A key advantage of DEA is that educational administrators or their nominated researchers can choose inputs and outputs to represent a particular perspective or approach. For example, key business drivers critical to success of the organization can be the outputs. Then, those variables that can be argued to manifest themselves as outputs become the inputs. A simple model of university efficiency might argue that when academic staff and buildings and grounds (inputs) are put together, they give rise to enrolments (output). Hence, a resource is classified as an input while anything that uses resources is classified as an output. DEA forces policy-makers to explicitly state the objectives of the organization. Ultimately, these objectives become the outputs in efficiency modeling and the resources needed become the inputs.

There is no definitive study to guide the selection of inputs/outputs in educational applications of DEA. While outputs can be generally categorized into teaching, research, and service, it is very difficult to find true measures for these dimensions (Ahn and Seiford 1993, p. 197). In short, it is possible for the analyst to select a parsimonious set of desired outputs, provided they can be reasoned to be manifestations of inputs. There is thus a pressing need for the choice of inputs and outputs to reflect the industry or the setting examined.

Table 15.1 Overall performance (model 1)[a]

Inputs	Outputs
Academic staff, FTE [AcaFTE]	Undergraduate enrolments, EFTSU [UgEFTSU]
Non-academic staff, FTE [NonAcaFTE]	Postgraduate enrolments, EFTSU [PgEFTSU]
	Research quantum [RQ]

[a]*Notes*: *FTE* full-time equivalence, *EFTSU* equivalent full-time student unit, measures student load. Research quantum is expressed as share of Commonwealth Government grants (%)

Accepted theories in different fields can also be employed to help select the inputs and outputs. In this study, production theory provides the starting point for efficiency modeling.

Production theory is concerned with relationships among the inputs and outputs of organizations (Johnes 1996). This approach requires the specification of inputs and outputs in quantitative terms. According to Lindsay (1982) and Johnes (1996), some of the generally agreed inputs of universities can be classified as human and physical capital, and outputs as arising from teaching and research activities. In selecting variables, controllable inputs and those outputs of particular interest to administrators are preferred. However, there is always the danger of excluding an important performance variable due to lack of suitable data or to limitations imposed by small sample sizes. Therefore, it is essential to develop a good understanding of the inputs and outputs as discussed previously before interpreting results of any efficiency model.

15.2.3 Research Design

Calculation of relative efficiency scores with different models generates insight into the performance of universities on various dimensions, thus guiding managerial action (Nunamaker 1985). The production approach was used here to design the first performance model called "overall performance of universities" (see Table 15.1). The argument in the overall performance model is that universities employ people to produce enrolments, and generate research output. It is assumed that staff numbers capture the physical capital input dimension. Inclusion of enrolments as output recognizes the importance of this measure in Federal funding. In this instance, research output is represented by Research Quantum, which was the research component of Federal funds given to universities in recognition of their share of overall research activity during the study period of 1995. Part of Research Quantum was indexed to the number of publications, which carried a 12.5% weighting in the Composite Index used to calculate the Research Quantum. (Other components of the index were competitive research grants earned and higher degree research completions; see DEETYA 1997).

The second performance model focuses on delivery of educational services (see Table 15.2). The inputs are the same as in the first model. The purpose here

Table 15.2 Performance on delivery of educational services (model 2)

Inputs	Outputs
Academic staff, FTE [AcaFTE]	Student retention rate (%) [RetRate]
Non-academic staff, FTE [NonAcaFTE]	Student progress rate (%) [ProgRate]
	Graduate full-time employment rate (%) [EmplRate]

Table 15.3 Performance on fee-paying enrolments (model 3)

Inputs	Outputs
Academic staff, FTE [AcaFTE]	Overseas fee-paying enrolments, EFTSU [OverEFTSU]
Non-academic staff, FTE [NonAcaFTE]	Non-overseas fee-paying postgraduate enrolments, EFTSU [FeePgEFTSU]

is to compare how successful staff are in delivering the educational services that contribute to enrolling graduating students into other courses in the same university (retention rate); successful student subject load (student progress rate); and, proportion of graduates in full-time employment as a percentage of graduates available for full-time work (graduate full-time employment rate).

The third performance model is designed to explore the success of universities in attracting fee-paying students. In the context of ongoing changes in the Australian tertiary education sector, there is a growing pressure on universities to develop this revenue source by increasing the number of fee-paying enrolments. The model depicted in Table 15.3 has the same inputs as in the previous two models but the outputs are replaced by overseas fee-paying enrolments and nonoverseas fee-paying postgraduate enrolments. For the period under study, there were no reliable figures available on the more recent practice of nonoverseas fee-paying undergraduate enrolments.

The choice of variables in the above performance models can be traced to literature. For example, staff numbers appear as inputs in Tomkins and Green (1988). Similarly, student numbers appear as outputs in Ahn (1987), Ahn et al. (1988, 1989), Ahn and Seiford (1993), Beasley (1990), Coelli et al. (1998), and Tomkins and Green (1988). Ahn (1987) and Ahn et al. (1988, 1989) also use research income from the Federal Government, that is, Research Quantum, as an output. Breu and Raab (1994) use student retention rate as an output. The remaining output variables in the performance models tested here represent dimensions that were also deemed important, such as graduate full-time employment rate and student progress rate.

Data for inputs and outputs identified in Tables 15.1–15.3 were extracted from publications by the Department of Employment, Education, Training and Youth Affairs (DEETYA) for 1995. Selected Higher Education Statistics series and the report by Andrews et al. (1998) were the main sources of data. The 36 universities investigated correspond to the total number of universities in the period and represent a healthy sample size (see Table 15.8 in Avkiran 2001).

Table 15.4 Potential improvements for universities with the lowest BCC scores

	Unit with lowest score and the potential improvements				
Performance model	Inputs (%)[a]			Outputs (%)	
Model 1 (overall performance)	Unit 34 (44.88%)	AcaFTE	↓ 16[b]	UgEFTSU	↑ 122[c]
				PgEFTSU	↑ 122[c]
				RQ	↑ 122[c]
Model 2 (performance on delivery of educational services)	Unit 18 (87.83%)	AcaFTE	↓ 30[b]	RetRate	↑ 17[c]
				ProgRate	↑ 13[c]
				EmplRate	↑ 39[c]
Model 3 (performance on fee-paying enrolments)	Unit 12 (15.32%)	AcaFTE	↓ 16[b]	OverEFTSU	↑ 552[c]
				FeePgEFTSU	↑ 798[c]

[a] These represent input slacks (over-utilized inputs) since DEA tests are run under output maximization
[b] ↓ Input can be lowered
[c] ↑ Output can be raised

15.2.4 Results and Analysis

For brevity, Table 15.4 shows the potential improvements only for the most inefficient universities in each performance model. In the case of unit 34 in model 1, the university can increase its enrolments of undergraduates and postgraduates as well as its research output by 122%. Furthermore, the above increases in outputs can be achieved while simultaneously reducing academic staff numbers. Indicated reduction in academic staff numbers is an input slack and represents an over-utilized resource. The total sample input slacks range between 1.34 and 5.26% (model 1), 19.73 and 36.51% (model 2), and negligible values for model 3. That is, overall, slacks are small. On the other hand, as suggested by the relatively low mean and high standard deviation of scores in model 3 (not shown here), there is, for example, a potential to improve outputs by up to eightfold in unit 12 (see Table 15.4).

It is also possible to implement a more detailed reference comparison. This involves comparing an inefficient unit's usage of inputs and production of outputs with those of one of the efficient units in its reference set (as part of benchmarking). When there is more than one efficient unit in the reference set, the unit that makes the largest contribution to computation of the inefficient unit's score can be selected (i.e., unit with the largest peer weight or lambda). In the case of model 1, there are four peers in the reference set of the inefficient unit 34. These peers and their corresponding weights are 2(0.073), 5(0.308), 7(0.195), and 8(0.424). Expanding this example, inefficient unit 34 is compared to efficient unit 8. Test results reveal that efficient unit 8 uses less inputs to generate more outputs (see Table 15.5).

The technical efficiency scores for model 1 decomposed into PTE and SE, and the nature of returns to scale for the complete sample can be seen in Avkiran (2001, Table 15.7). In summary, 13 universities were operating at MPSS, 4 at IRS, and the remaining 19 at DRS. With DRS, an increase in inputs leads to a less than proportionate increase in outputs (Banker 1984). This implies that, as per 1995 data,

Table 15.5 Reference comparison of input usage and output production between Unit 8 (efficient) and Unit 34 (inefficient) in model 1

Unit 8's usage of inputs relative to that of Unit 34	Unit 8's production of outputs relative to that of Unit 34
46% of AcaFTE	184% of UgEFTSU
62% of nonAcaFTE	122% of PgEFTSU
	105% of RQ

universities operating at DRS had grown larger than their MPSS and could consider downsizing, albeit by small amounts. Fourteen of the 19 DRS universities are above average in size, where the FTE academic staff is used as a proxy for size. On the other hand, the average size of the MPSS universities is about 852 FTE academic staff, which is substantially below the sector average of 1,237. Among the 23 scale-inefficient universities, 13 were also inefficient on pure technical efficiency. It should be noted that, in general, it is easier to reduce technical inefficiency than it is to reduce scale inefficiency. Thus, we anticipate management would be more interested in reducing technical inefficiency first before dealing with scale inefficiency.

15.2.5 Concluding Remarks

The main objective of this study was to apply DEA in examining the relative efficiency of universities. The greatest potential improvements were in raising outputs such as fee-paying postgraduate enrolments (a significant source of revenue for schools), reflecting the scope for further development in this key revenue-raising activity. There were also small slacks in input utilization, which could be addressed by university administrators without too much difficulty. Based on performance model 1, a small number of universities were operating at IRS, about one-third at MPSS, and the remaining majority at DRS. Eleven of the 19 DRS universities were also found technically inefficient. This suggests that managers are likely to focus first on removing the technical inefficiency of these universities before addressing the longer term and more demanding exercise of restructuring the scale of operations. Nevertheless, an average scale efficiency of 94.20% (model 1) suggests a small potential to downsize the sector.

Decomposing technical efficiency scores into pure technical efficiency and scale efficiency provides guidance on what can be achieved in the short- and long term. For example, if the majority of inefficiency is due to the small size of operations, that is, IRS, then that DMU will need to plan for expansion. While this exercise can be accelerated through in-market mergers in the case of profit-making organizations, it is a more cumbersome process for universities, which normally have to proceed through a lengthy consultation process with

Federal and State governments. On the other hand, pure technical inefficiency can usually be addressed in the short term without changing the scale of operations. The observation that 19 Australian universities were operating at DRS in 1995 may be an indication of counter-productive university amalgamations implemented in the early 1990s.

Collection and classification of research-related data merit further comment. While there may be some disagreement on the classification of research income as an input, it is theoretically appropriate since it represents a resource rather than a research outcome. Research outcome (output) can be in the form of a rating, or number of publications, or money indexed to number of publications that eventually finds its way back to the university. Of course, when the focus of efficiency analysis shifts from production of research to a purely financial outlook, then it is quite acceptable to list research income as an output.

There are also potential problems in collecting data on research income and research outcome. The long and variable lead times between receipt of research income and publications means one cannot easily put together corresponding data for a particular year. In Australia, an additional difficulty has been the unreliable nature of refereed publications reported by universities where independent audits by the consulting firm KPMG have, in the past, revealed mistakes in classification and counting of publications, casting doubt on whether research outcomes can be satisfactorily quantified.

As a final remark, while universities are more than ever evaluated on quantifiable inputs and outputs, higher education still retains certain key characteristics that set it apart from other types of organizations. These key characteristics are "the lack of profit motive, goal diversity and uncertainty, diffuse decision making, and poorly understood production technology" (Lindsay 1982, p. 176). Lindsay's comments maintain most of their currency today. It is this nature of higher education that continues to complicate selecting inputs/outputs and undertaking interinstitutional comparisons that satisfy all parties. This particular nature of performance measurement in higher education makes research in this field both challenging and rewarding.

15.3 Hotels[2]

15.3.1 Introduction

This section examines productivity of Queensland's largest hotels. The primary aim of the section is to illustrate the potential use of DEA in the hospitality industry. The focus is on the technical efficiency component of productivity.

[2]ADAPTED FROM AVKIRAN 2002.

The study distinguishes between pure technical efficiency and scale efficiency, while determining the nature of returns to scale for each unit. Explanation of key concepts and implementation of DEA in a plain language empower the hospitality professional as well students of productivity measurement.

The hospitality industry appears to be low on productivity where the service sector itself compares unfavorably to other sectors (Ball et al. 1986; Witt and Witt 1989; Johns and Wheeler 1991). This observation is sometimes attributed to the hands-on style of hotel managers who may not put much stock in formal planning (Guerrier and Lockwood 1988). Others have accused hospitality managers of not being proactive (Olsen et al. 1991). Yet, there is a pressing need to address productivity analysis in the hospitality industry if hotels are to exist as sustainable business entities in rapidly maturing markets. More to the point, focusing on resource utilization at the expense of revenue generation or vice versa leads to a limited analysis. Therefore, it is essential to consider inputs and outputs simultaneously in the quest for productivity improvement (Ball et al. 1986), which is a task conducive to DEA.

15.3.2 Conceptual Framework

Measuring productivity forces a business to clearly identify its objectives and goals. In essence, such an exercise provides direction and control. Yet, in practice there are problems defining and measuring the service outputs that may be mostly intangible. Renaghan (1981) suggests that hotel clients evaluate their experience as a whole rather than in distinguishable components, which further compounds the measurement problem. Briefly, some of the complicating factors in measuring productivity in the service sector include intangible variables, the need for production and delivery of service in real time (McLaughlin and Coffey 1992), and limitations of productivity ratios borrowed from the manufacturing sector.

It is also essential to be aware of the interaction between key variables. For example in the hospitality industry, up-market hotels that provide personalized high quality service to their guests normally employ more people per guest. On the other hand, reducing the number of employees in such hotels in the name of raising productivity is likely to lower the quality of customer service that their clientele expect. In general, the extent machines can substitute for a human is limited in the service sector (Mill 1989). This observation highlights the importance of productive employees.

Examples of productivity ratios used in hospitality industry include kitchen meals produced per number of kitchen staff, total guest rooms per total kilowatt hours, number of satisfied hotel customers per total number of hotel customers, and hotel revenue per total management salaries (Ball et al. 1986). Clearly some of these ratios measure physical inputs and outputs whereas others measure financial performance.

In a survey of Zimbabwean hotel managers, Messenger and Mugomeza (1995) report similar physical and financial productivity ratios. In addition, their findings indicate a preference among general managers for financial ratios. One thing the above productivity ratios have in common is that each ratio reflects a narrowly defined measure of productivity. Yeh (1996, p. 980) succinctly explains the principal shortcomings of traditional ratio analysis as, ". . .each single ratio must be compared with a benchmark ratio one at a time while one assumes that other factors are fixed and the benchmarks chosen are suitable for comparison." Therefore, there is a need for a technique that can bring key productivity ratios together to produce a simultaneous measurement of productivity with a wider scope, where units are benchmarked on observed best performances.

Productivity paradox in hospitality suggests that increasing physical output beyond certain levels without increasing resources is bound to have a negative impact on quality. On the other hand, quality is always a difficult concept to measure in the service sector; Haywood (1983) provides a good discussion of the dimensions of hospitality service and its assessment. The potential positive correlation between quality and productivity has also been recognized (Schroeder 1985; Butterfield 1987). In this instance, Schroeder's definition of quality includes the prevention of mistakes in delivery of services, which in turn helps to maintain higher productivity since resources are not wasted on remediation.

Another reason to investigate quality is the nature of the hospitality industry where price is no longer a key competitive tool and businesses tend to compete on service quality and image (Witt and Moutinho 1994). Unfortunately, detailed comparable data on hotel service quality are still hard to find. An alternative approach is to focus on sales revenue and costs, which lends it to comparisons between hotels. If productivity is defined as the ratio of sales revenue to costs, this provides a crude measure of a hotel's overall performance (Johns and Wheeler 1991).

Jones (1988) discusses a framework of inputs, intermediate output, output and outcomes for the hospitality industry as per Flynn (1986), where output and outcomes distinguish between consumption and satisfaction, respectively. He proposes this framework as a tool for analyzing the operational links between productivity, capacity, and quality. Interrelationships between the key factors of service industries are best summarized by a quote from Jones:

> . . . services vary from one service encounter to the other (heterogeneity) largely because they depend on the interaction of the consumer with the service provider (simultaneity), providing "something" that the user cannot easily objectively measure (intangibility), which makes it almost impossible for the service provider to store (perishability). (1988, p. 105)

Jones (1988) further posits that productivity improvements are most likely in the conversion of inputs to intermediate output, where the manager has more control. The example given states that a restaurant's kitchen staff use various commodities (inputs) to prepare food for a set menu based on expected orders

(intermediate output). Actual output may be less than the intermediate output when some food is not sold. On the other hand, outcomes experienced by customers may range from satisfying hunger to social contact.

15.3.3 A Quick Guide to Selecting Inputs and Outputs

A sensible approach for management is to identify the key business drivers (outputs) and the inputs that they can link to producing these outputs. Geller (1985) provides some guidance. As a result of interviewing 74 hotel executives, Geller determines the goals and factors considered critical to a hotel's success. Geller argues that a reliable set of critical success factors are impossible to develop unless there are well-defined goals first. Geller distinguishes critical success factors from the measures used to monitor them. In fact, key business drivers are closer in definition to what Geller defines as measures of critical success factors. A comprehensive list of 40 key performance indicators in hotels appear in Sandler (1982, pp. 158–159). To clarify the links between goals, critical success factors and measures, a couple of the examples from Geller (1985) are presented next.

For instance, to increase market share (a company goal) a hotel needs to have a better service (a critical success factor) than its immediate competitors. This service can be measured through, for example, ratio analysis of repeat business, occupancy rates, and informal or formal feedback. On the other hand, owner satisfaction (another company goal) would require attention to adequate cash flow (a critical success factor), which in turn can be measured through sales revenue, gross profit, and departmental profit. There is no universal set of critical success factors appropriate for every hotel. In fact, different hotels are likely to emphasize different critical success factors reflecting varying organizational structure, stage in life cycle, financial status, and so on. Thus, in practice, some variation is likely in the final choice of measures for critical success factors or key business drivers, that is, outputs, as well as choice of inputs.

15.3.4 A Numerical Example

Data on the largest hotels (ranked by number of guestrooms) were extracted from the periodical Business Queensland (1997). The three inputs are *full-time* staff numbers, *permanent part-time* staff numbers, and *total bed capacity* (i.e., number of beds). The two outputs are *revenue* and *cost of a double room* (i.e., rate charged to guests). It is argued that the management's ability to fill the beds is reflected in higher revenue. In addition, the staff's ability to distinguish their hotel from other similar hotels through better service is expected to manifest itself in a higher rate for a double room. These arguments suggest that the more appropriate choice

Table 15.6 Relative efficiency scores and returns to scale

DMU	Hotel	TE	PTE	SE	RTS
1	Hamilton Island Resort	49.89	100.00	49.89	DRS
2	Hotel Conrad and Jupiters Casino	100.00	100.00	100.00	MPSS
3	Ramada Hotel Surfers Paradise	89.02	91.78	96.99	DRS
4	ANA Hotel Gold Coast	83.44	88.85	93.91	DRS
5	Sea World Nara Resort	100.00	100.00	100.00	MPSS
6	Mercure Resort Surfers Paradise	72.56	73.60	98.59	CRS
7	Novotel Twin Waters Resort	25.67	28.78	89.20	DRS
8	Matson Plaza Hotel	96.90	99.68	97.21	CRS
9	Novotel Palm Cove Resort	34.59	70.24	49.25	CRS
10	Daydream Island Travelodge Resort	73.00	76.67	95.22	CRS
11	The Pan Pacific Hotel Gold Coast	100.00	100.00	100.00	MPSS
12	Novotel Brisbane	83.30	83.45	99.82	DRS
13	Capricorn International Resort	57.14	64.83	88.14	DRS
14	Cairns Hilton	89.98	91.36	98.49	CRS
15	The Heritage Hotel	100.00	100.00	100.00	MPSS
16	The Tradewinds Esplanade Hotel	95.07	97.04	97.97	CRS
17	Kingfisher Bay Resort & Village	65.04	69.95	92.98	CRS
18	Mercure Inn Townsville	90.12	100.00	90.12	IRS
19	Hotel Grand Chancellor Brisbane	100.00	100.00	100.00	MPSS
20	Sheraton Noosa Resort	100.00	100.00	100.00	MPSS
21	Club Crocodile Resort	100.00	100.00	100.00	MPSS
22	Laguna Quays Resort	100.00	100.00	100.00	MPSS
23	Club Crocodile Long Island	76.34	77.43	98.59	CRS

Notes: *TE* technical efficiency, *PTE* pure technical efficiency, *SE* scale efficiency, *RTS* returns to scale, *DRS* decreasing returns to scale, *MPSS* most productive scale size, *IRS* increasing returns to scale

for DEA in this case is output orientation, identifying the behavioral objective of the organization as *maximizing outputs with given inputs*. After deletion of cases with missing values, 23 Queensland hotels remained from the original sample of 50.

In this study, publicly available data on hotels are used. A downside of this approach is that only a limited set of inputs and outputs are readily available. For example, it would have been desirable to have a direct measure of customer service quality for the group of 23 Queensland hotels. Nevertheless, the aggregate measures used indirectly measure the service quality at these establishments. The Queensland hotel sample was analyzed with output-oriented BCC model. Thirteen of the hotels are inefficient, indicating an acceptable level of discrimination. Technical efficiency is decomposed into pure technical efficiency and scale efficiency, where scale efficiency is the component of technical efficiency that can be attributed to the size of operations. Thus, scale inefficiency represents deviations from the MPSS, an optimal size identified when the CCR and BCC scores both equal one.

Table 15.6 indicates that eight of the ten efficient hotels are, in fact, operating at the MPSS. However, some of the hotels are also flagged for a potential reduction in

Table 15.7 Percent potential improvements for the three most inefficient hotels

Inputs			Outputs	
Full-time staff	Part-time staff	Bed capacity (%)	Revenue (%)	Room rate (%)
Novotel Twin Waters Resort (28.78%)				
Peers: 2, 5, 15, **20**				
0	0	−30.66	247.46	247.46
Capricorn International Resort (64.83%)				
Peers: 2, **5**, 15, 20				
0	0	−16.43	54.25	54.25
Kingfisher Bay Resort & Village (69.95%)				
Peers: 2, 5, **20**				
0	−45.42%	−22.96	42.96	42.95

Notes: Pure technical efficiency of the hotel is in brackets. The peer to emulate is in bold

bed capacity – an over-utilized input. Potential input reductions during output maximization are known as input slacks. In this sample, 12 of the hotels should be able to raise their revenue and room rates while reducing one or more of their inputs (all else equal). However, higher room rates are likely to be realized only after improvements in service.

Potential improvements for the three inefficient hotels are shown in Table 15.7, accompanied by peers (reference sets). The global leader, that is, the overall best performer or the most often benchmarked efficient unit, is Sea World Nara Resort (12 times). In the absence of a more detailed analysis, management of an inefficient hotel can benchmark the operations of Sea World Nara Resort. However, knowledge of peers in the reference set provides a greater wealth of information. In this instance, management is able to examine these efficient peer hotels and compare their resource allocations and outputs. A more focused reference comparison involves comparing an inefficient unit's usage of inputs and production of outputs with those of one of the efficient units in its reference set. When there is more than one efficient unit in the reference set, the unit that makes the largest contribution to computation of the inefficient unit's score can be selected (i.e., unit with the largest peer weight or lambda). Such peers are depicted in bold in Table 15.7.

15.3.5 Concluding Remarks

Findings from the cross-sectional data suggest that some of the hotels can significantly reduce their bed numbers and part-time staff numbers while increasing outputs of revenue and room rates. The potential to charge a higher rate for a double room should be interpreted in the context of the inefficient hotel. For example,

management ought to take into account the characteristics of the local market including the presence of competitors and the type of guests that are normally attracted to the area. It is also essential to consider how to improve the service. A cost-benefit analysis should accompany such an investigation. This will ensure that any additional expenses incurred in service improvement are actually fully recovered by higher rates charged to guests.

Similarly, implementation of the recommended simultaneous reduction in staff numbers should follow an initial examination of input output combinations employed by efficient peers. This is an essential part of the benchmarking exercise that needs careful execution. Particularly in high-class hotels where guests expect personal service, the number of staff that can be cut without undermining service quality is limited.

Organizations can benefit in a number of ways if management adopt DEA as one of their tools for analyzing operational productivity. The first point to make is that DEA is often used in conjunction with other benchmarking processes already in place, as well as such techniques as the balanced scorecard for gauging an organization's performance (e.g., see Denton and White 2000). The nonprescriptive nature of DEA allows management to develop their own productivity or performance models that best reflect their business aspirations.

DEA enables management to integrate dissimilar multiple inputs and outputs, for example, accounting measures and customer service quality, thus making simultaneous comparisons that would otherwise not be possible. Inputs or outputs indicated by DEA for potential improvements should encourage management to explore better ways of running their business. On an everyday basis, management can use DEA to test their established knowledge of operations and initiate investigations when contradictions arise. Often, the process of DEA forces management to clearly identify the links between business goals and key business drivers, essentially providing direction and control.

15.4 Real Estate Agents[3]

15.4.1 Introduction

In this section, we follow through an application of DEA in the field of property management, where we demonstrate the use of weight restrictions. The data consist of variables that were reported in a profitability study of the residential real estate agency practice in Queensland (REIQ 1994). The overall aim is to compare the conclusions from a traditional accounting-based performance measure with that of DEA.

[3]ADAPTED FROM AVKIRAN 2006.

Since mid-1990s the residential real estate practice has been characterized by a low-turnover market despite the all-time low home loan interest rates. This is mostly attributed to slow economic growth and low inflation rates in Australia, coupled with moderate unemployment rates. The lack of consumer confidence has stymied the turnover in the residential real estate market, leading to narrower profit margins for most of the agents. This downward pressure on the revenue side has, in turn, refocused the attention of agents on finding savings on the expense side of their business. The real estate agency practice is currently conducive to a search for cost savings. This search is particularly suitable through DEA under the analysis option of input minimization.

Often, performance of real estate agents is measured by accounting variables found on the profit and loss statement. While this time-honored approach allows each agent to compare its operating surplus to that of others, it does not easily lend itself to either multidimensional benchmarking or understanding the interplay between multiple inputs and outputs. In this section, we reexamine an extract of the data behind the Profitability Report 1992/93 prepared by the Real Estate Institute of Queensland, 1994 (REIQ). Revenue and expense data were collected based on accrual accounting.

15.4.2 Research Design

The research design aims to explore the correspondence between the results of DEA and the traditional profitability analysis. Initial exploration of the raw data reveals that the variables fees, recoverable, and travel have too many missing values. These variables are removed. When we calculate the range for each variable, we identify those with ranges in excess of five decimal places.[4] Sorting the DMUs in descending order helps to determine the units at the two extremes. After deletion of such DMUs and a small number of units with missing values, we are left with 87 units for DEA. Once the DEA is run, we look to see whether those agents that are top performers in the profitability measurement (i.e., operating surplus) are also appearing as efficient units under DEA. First, let us address the selection of inputs and outputs from the available pool of variables.

In general, we regard personnel and expense categories as inputs (resources) and revenue items as outputs. Looking at Table 15.8 we see that there are three outputs which we should include in the productivity model. On the input side, we collapse the personnel items into a single variable, "staff numbers." Similarly, bank charges, insurance, interest, personnel, postage, printing, professional fees, subscriptions, and telephone are aggregated as "other operating expenses." Salaries of employees

[4] DEA can give meaningless results if data on a variable have a very large range. This is because DEA is an efficient frontier technique and a DMU with an extreme value on a variable can appear as efficient even though it may be performing poorly on all other variables.

Table 15.8 REIQ questionnaire items used to collect profitability data

Revenue	Sales	
	Property management	
	Other	
Expenses	Advertising, promotion	Printing, stationery
	Bank charges	Premises
	Equipment, plant	Professional fees
	Group fees (omitted)	Recoverable (omitted)
	Insurance	Salaries – employees
	Interest	Subscriptions – donations
	Motor vehicles	Telephone, fax
	Personnel	Travel, parking (omitted)
	Postage, couriers	
Personnel	Principals	
	Sales persons	
	Property managers	
	Administration/clerical	

Table 15.9 Inputs and outputs in the real estate productivity model

Inputs (controllable)	*Outputs*
Staff numbers	*Sales*
Advertising, promotion	*Property management*
Equipment, plant	*Other*
Motor vehicles	
Premises	
Other operating expenses	

are omitted since staff number is already an input. The emerging input/output list is depicted in Table 15.9.

Weight restrictions warrant special comment before we examine the findings. The manager needs to exercise some value judgment before introducing restrictions on the weights assigned by the optimization process. Nevertheless, this value judgment need not take place in an empirical void. Normally, the first step is to examine the weights attached to inputs and outputs in an unrestricted run. If we find that certain variables have been under- or over-represented in the calculation of efficiency scores, we then seek a consensus about the relative importance of outputs and inputs among those who are best familiar with the interactions of the variables modeled (Wong and Beasley 1990).

In the exercise demonstrated in this section, the consensus was reached after consultations with colleagues who teach in this field (in the workplace, the consensus will be reached among managers). It was agreed that staff numbers represent the most important resource (input) and should not be allowed to fall below a weight of 30%. This contention is supported by the correlation between sales dollars and staff numbers at 0.80. This decision is also consistent with the observation that staff salaries form the largest expense item for the agents and that we cannot ignore this item when searching for cost efficiencies.

Imposing a lower bound of 30% on staff numbers ensures a more sensible allocation of weights during the optimization process that determines the efficiency scores. In the absence of such a weight restriction, DMUs that have attracted a 100% weighting on certain variables (and, 0% weighting on others) may unreasonably appear as efficient. That is, some DMUs may be evaluated as using very small amounts of inputs to generate outputs. This is a potential problem with total weight flexibility found in an unrestricted DEA run. Optimization can allocate such low weights to certain outputs or inputs that they are effectively omitted from calculation of efficiency scores (Dyson and Thanassoulis 1988).

15.4.3 Analysis of Results

Finding evidence for the need to impose weight restrictions on certain variables is an involved process. It requires examination of input/output contributions to calculation of efficiency scores to see if certain variables are consistently under-or over-represented. A range based on consensus can then be applied on a variable or variables. In this study, staff numbers make no contribution to calculation of efficiency scores in 27 of the DMUs. This is an undesirable situation and it provides empirical support for applying a lower weight restriction of 30% as agreed. A comparison of efficiency scores between the unrestricted model and the weighted model using the CCR (Charnes et al. 1978) follows.

All the units that are listed as efficient in the weight-restricted run are also efficient in the unrestricted run. In the weight-restricted model, nine of the previously efficient units become DEA inefficient. Now these inefficient nine units have scores ranging from 98.62 to 68.04%. It appears that limiting the lower weight of staff numbers adds more discriminatory power to the analysis. Specifically, units that were appearing as efficient because of the omission of staff numbers in weighting of scores are now reassessed as inefficient.

But, are the efficient units in the weight-restricted model the top performers as measured by operating surplus? Table 15.10 provides the answer. The five most profitable agents are also listed as efficient under DEA. However, as we go down the list the results are mixed. Nevertheless, it is possible to observe a general trend for higher efficiency scores to correspond to higher operating surpluses and lower efficiency scores to congregate with the lower operating surpluses. In fact, the bivariate correlation between operating surplus and efficiency is 0.57. The less than perfect correlation suggests that some of the high performers (on operating surplus) are not necessarily operating at optimal efficiency when compared to their peers. Let us first closely examine the case of agent 185 which is ranked 6th in a sample of 87 units on operating surplus, yet appears as only 82.58% efficient (see Table 15.10).

The potential improvements for agent 185 are reported in Table 15.11. Briefly, there are no output slacks but all the inputs can benefit from a reduction (i.e., cost savings). The largest potential saving is indicated for reducing expenditure on

Table 15.10 Comparing agents sorted on operating surplus in descending order

Agent ID	Operating surplus ($)	Efficiency score (%)	Agent ID	Operating surplus ($)	Efficiency score (%)
204	1,484,776	100.00	116	286,815	84.16
201	1,351,387	100.00	113	273,244	100.00
196	934,377	100.00	166	269,912	83.76
195	846,151	100.00	111	268,634	80.46
189	799,600	100.00	104	257,520	72.78
185	**716,068**	**82.58**	115	245,612	55.67
186	700,738	78.11	114	244,836	75.60
179	695,100	100.00	124	241,266	93.65
171	684,988	100.00	117	239,913	77.54
183	677,794	95.40	99	228,052	68.04
180	675,235	98.62	102	228,036	77.85
188	651,750	86.46	103	228,036	83.94
172	644,445	100.00	106	223,000	72.82
169	638,045	100.00	126	222,500	70.77
177	588,767	90.12	78	219,483	100.00
173	586,491	76.63	91	215,784	100.00
164	583,500	100.00	96	215,093	61.75
165	570,398	94.82	110	210,928	62.77
174	567,184	100.00	**81**	**202,361**	**100.00**
167	553,820	87.91	79	201,281	79.72
175	538,333	80.85	89	190,129	64.39
176	528,264	79.96	94	186,428	65.34
168	527,746	84.06	90	174,994	81.85
159	497,594	71.26	77	174,953	62.99
178	494,244	82.10	63	174,000	64.79
158	447,969	65.62	69	172,078	91.76
162	434,272	80.61	92	157,410	73.14
150	413,207	100.00	83	155,939	67.75
149	399,594	100.00	68	155,248	69.49
140	386,350	100.00	70	142,318	71.79
157	362,323	100.00	57	141,000	70.89
148	361,398	88.30	58	139,380	56.92
142	361,027	100.00	62	128,100	57.03
143	358,291	74.84	60	127,400	52.46
136	354,910	100.00	72	126,181	59.10
141	343,191	71.23	56	115,531	87.81
134	340,494	81.05	48	113,376	67.79
138	339,273	80.42	64	111,668	87.14
127	327,931	90.75	47	104,841	58.65
139	313,200	78.10	39	103,800	73.22
125	308,092	100.00	45	101,130	50.80
132	308,000	94.96	44	100,338	35.42
133	298,523	74.15	46	81,254	57.25
122	293,000	86.39			

Table 15.11 Potential improvements for agent 185

		Actual	Target	Potential improvement (%)
Inputs	Staff numbers	17	14.04	−17.42
	Advertising, promotion	60,709	50,133.74	−17.42
	Equipment, plant	14,244	9,992.15	−29.85
	Motor vehicles	20,462	16,897.6	−17.42
	Premises	34,254	28,287.09	−17.42
	Other operating expenses	103,054	74,324.22	−27.88
Outputs	*Sales*	*676,363*	*676,363*	*0*
	Property management	*199,667*	*199,667*	*0*
	Other	*72,761*	*72,761*	*0*

Efficient agents in the reference set of the inefficient agent 185 (i.e., its peer group): 196, 172, 171, 140, 136, 78
Global leader: agent 164 with a reference set frequency of 42

equipment, followed by "other operating expenses." However, since "other operating expenses" is a composite variable of nine other variables, it would be worthwhile for the manager to run a second DEA, where inputs in the current model are replaced by nine operating expense variables. In this way, it would be possible to pinpoint the inefficiency in "other operating expenses" with more accuracy.

The efficient peers of agent 185 that are comparable in their configuration of inputs/outputs are also listed in the footnote to Table 15.11. For the purpose of benchmarking, agent 185 ought to probe the operations of agents in its reference set. An examination of the business mix of the reference set can provide insight to inefficiencies. Yet, other units may well be operating efficiently but are only generating low operating surpluses. A case in point is agent 81, which is ranked 63rd on operating surplus but appears as 100% efficient under DEA.

15.4.4 Concluding Remarks

Information in Table 15.11 indicates that agent 185 can reach full efficiency if it can implement the indicated potential improvements. This will also have the effect of raising its operating surplus even further since we are proposing to reduce its expenses (inputs). Similar potential improvements can be identified for other inefficient agents. However, we should bear in mind that DEA is a relative measure that calculates the feasible performance based on what is observed in the sample. That is, DEA does not provide an absolute measure of efficiency. Another way of looking at this is to realize that other agents outside the sample may well be more efficient than the best we have in the sample.

Agent 185 is a profitable operation with suboptimal efficiency. This can be attributed to favorable environmental conditions in its business catchment area, such as small amount of competition. On the other hand, agent 81 is an efficient operation with suboptimal profitability. In this case, the agent is probably

confronting fierce competition or an inactive local market that is obstructing it from applying its efficient operations to generate a higher surplus.

We can also conceive a longitudinal study of residential real estate agents. Each period's data on the agent can be treated as a DMU where a common efficient frontier is set up. In such a design, efficiency scores would indicate the performance changes in response to changes over time in management strategies or technology (Gillen and Lall 1997).

We examined profitability data on a group of residential real estate agents using a weight-restricted DEA run. The results indicate that the accounting measure operating surplus corresponds to the relative efficiency score at the two extremes. However, this relationship is not so obvious when we look at the middle performers. There are cases that imply that efficient agents are not necessarily profitable operations and vice versa. There may well be environmental factors not captured in the model that are causing the discrepancies.

15.5 Commercial Banks[5]

15.5.1 Introduction

The main purpose of this section is to illustrate the under-utilized *network data envelopment analysis* (NDEA) in the context of banking. It is well known that traditional DEA does not provide sufficient detail for management to identify the specific sources of inefficiency embedded in interactions among various divisions that comprise the organization. Fortunately, NDEA provides the investigator with a method to pinpoint whether the overall inefficiency observed in a DMU resides in its head office or in one of its other functions. Theoretically, each organizational function and its subcategory can be treated as a sub-DMU. For example, subcategories under the head office of a bank can be treated as sub-DMUs whose intermediate outputs become inputs for the front office sub-DMUs. Similarly, the intermediate outputs from the front office sub-DMUs become inputs for the back office sub-DMUs. Färe and Grosskopf (2000) define an intermediate product as an item that is an output from a sub-DMU that becomes an input for another sub-DMU in the network.

Thus, unlike traditional DEA, network DEA can provide insight to the specific sources of organizational process inefficiency and enable management to devise targeted remedial action. In other words, NDEA permits a fuller access to the underlying diagnostic information that would otherwise remain outside management's reach in what Färe and Grosskopf (1996) refer to as the "black box." Nevertheless, application of NDEA in the context of the functional

[5]ADAPTED FROM AVKIRAN 2009A.

framework of a bank is not practical unless the exercise is undertaken by someone who has full access to internal managerial accounting records. Formulating the numerous interactions among the multiple functions and their corresponding subfunctions would require an extensive internal audit of business processes and development of a transfer pricing framework based on the principles of responsibility accounting (responsibility accounting requires costs and revenues to be traced to those divisions that are responsible for them; see Solomons 1968). In recognition of this practical problem, Sect. 15.2.2 revisits the process of identifying sub-DMUs.

Briefly, this section of the chapter identifies profit centers (sub-DMUs) and their corresponding expenses and revenues for the purpose of illustrating network DEA. In the empirical illustration, a DMU is a domestic commercial bank in the United Arab Emirates (UAE) and sub-DMUs are key profit centers therein. In so doing, we analyze the profit efficiency of a bank from a managerial accounting perspective, rather than focusing on physical or functional divisions. Initially, the study assesses the profit efficiency of domestic commercial banks in UAE for the fiscal year ending December 31, 2005 using traditional DEA in the form of slacks-based measure (SBM) (see chapter 8, Tone 2001, and Morita et al. 2005). Then the section demonstrates network SBM based on simulated profit center data.

15.5.2 Conceptual Framework

15.5.2.1 Modeling Profit Efficiency

In banking, a DMU is often described as either a bank or a bank branch, although the investigator can conceivably build a sample by collecting input–output data on any identifiable organizational unit (physical or otherwise). While DEA's ability to capture the interaction among multiple inputs and multiple outputs is its distinctive advantage over traditional ratio analysis, it suffers from a certain loss of information when data are aggregated at bank level. For example, when noninterest income (fee income) generated in a bank is used as one of the outputs in DEA, it is impossible to tell which of the profit centers is the main source of inefficiency. Such an investigation would normally require going beyond traditional DEA and undertaking additional examination of operations as part of an internal audit exercise. NDEA can assist in this type of in-depth organizational network investigation.

The study follows the common assumption that the banks' organizational performance can be investigated under the bank behavior model of intermediation (also known as the asset approach after Sealey and Lindley 1977), where deposits are converted into loans. This intermediation process can be summarily captured by the proxy inputs of *interest expense* (x_{ie}) and *noninterest expense* (x_{nie}), and outputs of *interest income* (y_{ii}) and *noninterest income* (y_{nii}). This approach

Table 15.12 Gross credit extended by UAE banks (x000,000 AED)

Credit categories or profit centers	2005 Total	Percent of grand total
Loans, advances and overdrafts (LAO)	361,564	91.56
Mortgaged real estate loans (MREL)	17,518	4.44
Discounted commercial bills (DCB)	15,811	4.00
Grand total bank credit	394,893	100.00

AED United Arab Emirates Dirham

effectively measures the banks' profit efficiency because the variables are costs and revenues as per banks' profit and loss statements. Consistent with the approach recommended by Dyson et al. (2001), this parsimonious input–output set is considered appropriate for covering the full range of resources used and outputs created, while providing adequate discriminatory power. Examples of other studies where this selection of variables have been used include Charnes et al. (1990), Yue (1992), Miller and Noulas (1996), Bhattacharyya et al. (1997), Brockett et al. (1997), Leightner and Lovell (1998), Avkiran (1999, 2000, 2009b), and Sturm and Williams (2004).

15.5.2.2 Key Profit Centers and Estimating Corresponding Data

Based on the annual report from the Central Bank of UAE (2006), Table 15.12 summarizes the principal credit categories or profit centers. Proportions corresponding to the profit centers later on become weights in Network slacks-based measure (NSBM) of efficiency acting as proxies for the importance of various lending business generated by the average UAE bank.

Figure 15.1 (adapted from Figure 4 in Färe and Grosskopf 2000) shows the links among the three main profit centers by identifying the exogenous inputs, intermediate products, and final outputs in a five-node production network. Exogenous inputs such as risk management, human resources, information technology, and general administration would normally comprise the key *noninterest expenses*. Cost of funds, which would be valued by internal funds transfer pricing (FTP), becomes the key *interest expense* for the profit centers LAO (loans, advances, and overdrafts), MREL (mortgaged real estate loans), and DCB (discounted commercial bills). Final outputs are comprised of sales of products and services in the current period, captured by the *interest income* and *noninterest income* flows generated by such sales.

Opening the "black box" to expose the network of sub-DMUs, can reveal, for example, instances of cross-selling where business people opening an overdraft facility with the bank are encouraged to take out a mortgage as well. Such referrals represent intermediate outputs from the profit center 1 that become intermediate inputs to the profit center 3. Bank's internal activity-based costing systems (ABC) would recognize such referrals as noninterest revenue for the profit center 1, and noninterest expense for the profit center 3 for setting up and

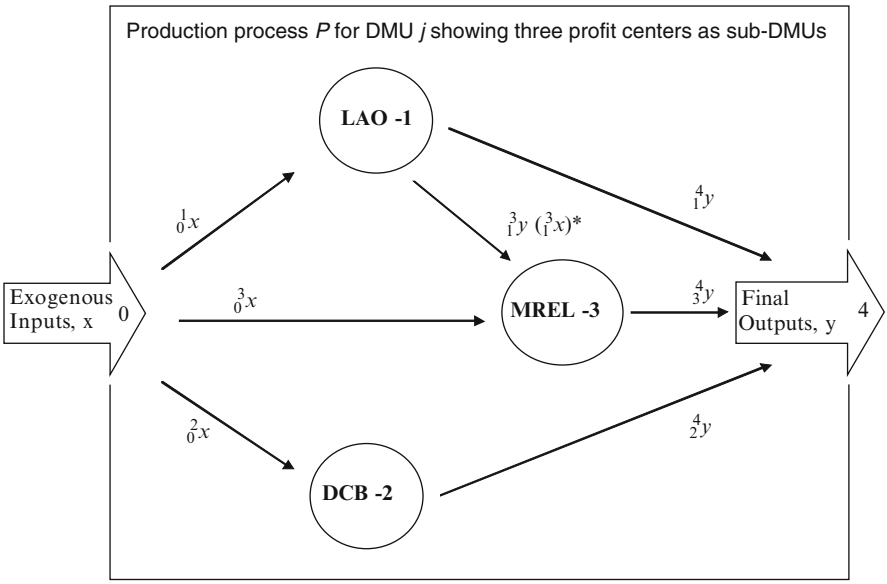


Fig. 15.1 Opening the "Black Box": network sub-DMUs. *Notes*: Inputs are denoted by *x* and outputs are denoted by *y*, where the *subscript* number identifies the origin and the *superscript* identifies the destination. *Asterisks* indicates an intermediate output from a sub-DMU that becomes an intermediate input for another sub-DMU. *LAO* loans, advances, and overdrafts, *MREL* mortgaged real estate loans, *DCB* discounted commercial bills

Table 15.13 Details of inputs and outputs for the profit centers (sub-DMUs)

LAO (1)		MREL (3)		DCB (2)	
Inputs (interest and non-interest expenses)	Outputs (interest and non-interest income)	Inputs (interest and non-interest expenses)	Outputs (interest and non-interest income)	Inputs (interest and non-interest expenses)	Outputs (interest and non-interest income)
$^{1}_{0}x_{ie}$	$^{4}_{1}y_{ii}$	$^{3}_{0}x_{ie}$	$^{4}_{3}y_{ii}$	$^{2}_{0}x_{ie}$	$^{4}_{2}y_{ii}$
$^{1}_{0}x_{nie}$	$^{4}_{1}y_{nii}$	$^{3}_{0}x_{nie}$	$^{4}_{3}y_{nii}$	$^{2}_{0}x_{nie}$	$^{4}_{2}y_{nii}$
	$^{3}_{1}y_{nii}{}^{a}$	$^{3}_{1}x_{nie}{}^{b}$			

Subscript numbers indicate the origin for a variable and superscripts indicate the destination
LAO loans, advances, and overdrafts, *MREL* mortgaged real estate loans, *DCB* discounted commercial bills
[a] Intermediate output
[b] Intermediate input resulting from intermediate output; interest expense (ie); noninterest expense (nie); interest income (ii); noninterest income (nii)

maintaining the accounts. In this study, NSBM measures the network production system, where the intermediate outputs and the intermediate inputs are treated as discretionary variables representing noncore profit center activities. According to Färe et al. (2007), Fig. 15.1, which represents a static model, is helpful in investigating the impact of intermediate products. The various inputs and outputs that flow from Fig. 15.1 are summarized in Table 15.13 for each profit center.

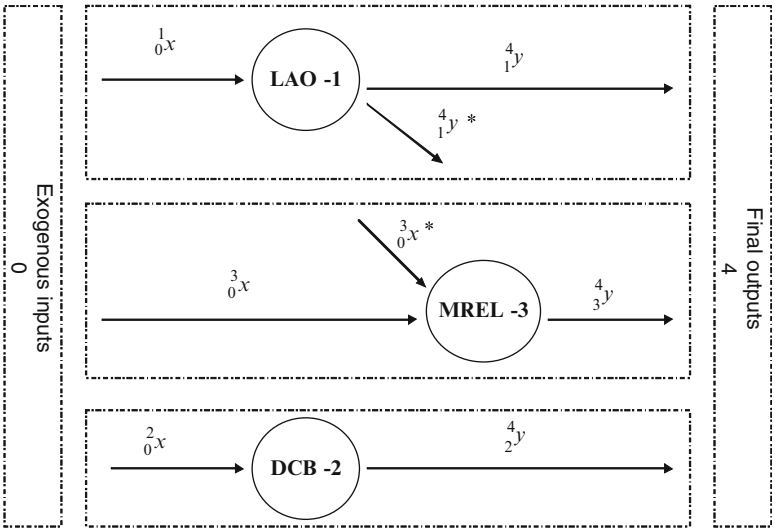


Fig. 15.2 Separation model. *Notes*: Inputs are denoted by x and outputs are denoted by y, where the *subscript* number identifies the origin and the *superscript* identifies the destination. *Asterisks* indicate the previously intermediate products. *LAO* loans, advances, and overdrafts, *MREL* mortgaged real estate loans, *DCB* discounted commercial bills

The separation model in Fig. 15.2 highlights what happens to the organization's production process if interactions between divisions are ignored. In short, the intermediate input in Fig. 15.1 now becomes an exogenous input for MREL and the intermediate output becomes a final output for LAO. Figure 15.2 depicts the three profit centers as independent production processes.

15.5.3 Methodology

15.5.3.1 Overview of Network DEA

We now return to the static network model depicted in Fig. 15.1 to formalize the mathematical relationships. In this section of the study, we closely follow the original notation in Färe and Grosskopf (2000). Under their network DEA, inputs, and outputs corresponding to various DMUs in the sample are required to satisfy the following properties:

$$x^{kn} \geq 0, \quad y^{km} \geq 0, \quad k = 1, \ldots, K, \quad n = 1, \ldots, N, \quad m = 1, \ldots, M \qquad (15.1)$$

i.e., non - negative inputs and outputs.

$$\sum_{k=1}^{K} x^{kn} \succ 0, \quad n = 1, \ldots, N \tag{15.2}$$

i.e., for every input, there is at least one positive value in the sample.

$$\sum_{k=1}^{N} x^{kn} \succ 0, \quad k = 1, \ldots, K \tag{15.3}$$

i.e., for every DMU, at least one input has a positive value.

$$\sum_{k=1}^{K} y^{km} \succ 0, \quad m = 1, \ldots, M \tag{15.4}$$

i.e., for every output, there is at least one positive value in the sample.

$$\sum_{k=1}^{M} y^{km} \succ 0, \quad k = 1, \ldots, K \tag{15.5}$$

i.e., for every DMU, at least one output has a positive value.
where there are $k = 1, \ldots, K$ DMUs (banks), $n = 1, \ldots, N$ inputs, and $m = 1, \ldots, M$ outputs. Thus, observations of inputs and outputs for each DMU can be summarized by the vector $(x^{kn}, y^{km}) = (x^{k1}, \ldots, x^{kN}, y^{k1}, \ldots, y^{kM})$. Explanations for the above variable properties (15.1–15.5) can be summarized as follows: (a) input and output values can be zero or positive; (b) there is at least one activity (DMU) where a particular input is used or an output is generated; and (c) each activity uses at least one input and produces at least one output.

Figure 15.1 shows a network of sub-DMUs called profit centers. A source of exogenous inputs (node 0) supplies the profit centers (nodes 1–3) whose final outputs are collected in a sink (node 4). Not all the available exogenous inputs are necessarily used up by the profit centers. Non-negative intensity variables z^k, $k = 1, \ldots, K$, are used to capture the extent a particular profit center participates in the production process. The series of equations below shows that network DEA can be envisaged as a group of models that share a common characteristic of having linear constraints (where subscript and superscript numbers, respectively, identify the origin and destination of variables) (Färe and Grosskopf 2000).

Allocation of exogenous inputs:

$${}^{1}_{0}x^{n} + {}^{2}_{0}x^{n} + {}^{3}_{0}x^{n} \leq x^{n}, \quad n = 1, \ldots, N \tag{15.6a}$$

Profit center 1:

$${}^{4}_{1}y^{m} \leq \sum_{k=1}^{K} z^{14}_{k1} y^{km}, \quad m = 1, \ldots, M^{1}, \quad \text{final output,} \tag{15.6b}$$

$$\sum_{k=1}^{K} z_{k0}^{11} x^{kn} \leq {}_{0}^{1}x^{n}, \quad n = 1, \ldots, N, \text{ exogenous input,} \tag{15.6c}$$

$${}_{1}^{3}y^{m} \leq \sum_{k=1}^{K} z_{k1}^{13} y^{km}, \quad m = 1, \ldots, M^{1}, \text{ intermediate output,} \tag{15.6d}$$

$$z_{k}^{1} \geq 0, \quad k = 1, \ldots, K \tag{15.6e}$$

Profit center 2:

$${}_{2}^{4}y^{m} \leq \sum_{k=1}^{K} z_{k2}^{24} y^{km}, \quad m = 1, \ldots, M^{2}, \text{ final output,} \tag{15.6f}$$

$$\sum_{k=1}^{K} z_{k0}^{22} x^{kn} \leq {}_{0}^{2}x^{n}, \quad n = 1, \ldots, N, \text{ exogenous input,} \tag{15.6g}$$

$$z_{k}^{2} \geq 0, \quad k = 1, \ldots, K \tag{15.6h}$$

Profit center 3:

$${}_{3}^{4}y^{m} \leq \sum_{k=1}^{K} z_{k3}^{34} y^{km}, \quad m = 1, \ldots, M^{3}, \text{ final output,} \tag{15.6i}$$

$$\sum_{k=1}^{K} z_{k0}^{33} x^{kn} \leq {}_{0}^{3}x^{n}, \quad n = 1, \ldots, N, \text{ exogenous input,} \tag{15.6j}$$

$$\sum_{k=1}^{K} z_{k1}^{33} x^{kn} \leq {}_{1}^{3}x^{n}, \quad n = 1, \ldots, N^{1}, \text{ intermediate input,} \tag{15.6k}$$

$$z_{k}^{3} \geq 0, \quad k = 1, \ldots, K \tag{15.6l}$$

Final outputs:

$${}_{1}^{4}y^{m} + {}_{2}^{4}y^{m} + {}_{3}^{4}y^{m} \leq y^{n}, \quad m = 1, \ldots, M. \tag{15.6m}$$

The above network allows for explicit modeling of intermediate products. A more in-depth exposition of general network DEA theory can be found in Färe and Grosskopf (1996, 2000) and Färe et al. (2007). The network SBM is explained next where the arguments in favor of NSBM are also presented.

15.5.3.2 Network Slacks-Based Measure of Efficiency

The study uses DEA-Solver Pro (SAITECH) software to execute weighted NSBM assuming variable-returns-to-scale and nonorientation. The objective function for the DMU and its respective constraints for divisional weights and intensity values are shown in (15.7) adapted from Tone and Tsutsui (2009).

$$\rho_o^* = \min \frac{\sum_{j=1}^{J} w^j \left[1 - \frac{1}{N_j}\left(\sum_{n=1}^{N_j} \frac{s_{no}^{j-}}{x_{no}^j}\right)\right]}{\sum_{j=1}^{J} w^j \left[1 + \frac{1}{M_j}\left(\sum_{m=1}^{M_j} \frac{s_{mo}^{j+}}{y_{mo}^j}\right)\right]} \tag{15.7}$$

subject to

$$\sum_{j=1}^{J} w^j = 1, \quad w^j \geq 0(\forall j), \tag{15.7a}$$

$$\sum_{k=1}^{K} z_k^j = 1(\forall j), \quad z_k^j \geq 0(\forall k, j), \text{ and} \tag{15.7b}$$

$$t^{(j,h)} z^h = t^{(j,h)} z^j \cdot (\forall (j,h)), \quad t_k{}^{(j,h)} = (t_1{}^{(j,h)}, \ldots, t_K{}^{(j,h)}) \in R^{T^{(j,h)} xK} \tag{15.7c}$$

where

k = a DMU (K = number of DMUs);
j = a division (J = number of division);
n = an input (N = number of input);
m = an output (M = number of output);
w = divisional weight;
s_{no}^- = input slack;
s_{mo}^+ = output slack;
Z = intensity;
(j, h) = intermediate product link between division j and division h;
$t^{(j,h)}$ = intermediate product;
$T^{(j,h)}$ = number of intermediate products.

Briefly, constraint (15.7a) indicates that weights are nonnegative for all the divisions and add up to 1; that is, we separately account for the importance of all the divisions. Constraint (15.7b) indicates nonnegative divisional intensities that add up to 1 (i.e., variable returns to scale). The last constraint, (15.7c), implies free linking where linking activities are discretionary while maintaining continuity between inputs and outputs (Tone and Tsutsui 2009); that is, an intermediate output from one division can become an intermediate input for another division.

The objective function for divisional efficiency is shown in (15.8). Optimal input and output slacks to emerge from (15.8) are entered into (15.7). Thus, the overall efficiency of a DMU captured in (15.7) is the weighted sum of divisional efficiency scores, but this overall efficiency score is neither the arithmetic

nor the harmonic mean of divisional efficiencies (Tone and Tsutsui 2009). Equations (15.7) and (15.8) are invariant to the units of measure used for inputs and outputs.

$$\rho_j = \frac{1 - \frac{1}{N_j}\left(\sum_{n=1}^{N_j} \frac{s_{no}^{j-*}}{x_{no}^{j}}\right)}{1 + \frac{1}{M_j}\left(\sum_{m=1}^{M_j} \frac{s_{mo}^{j+*}}{y_{mo}^{j}}\right)}, \quad (j = 1, \ldots, J) \tag{15.8}$$

Nonorientation is employed because it can accommodate the simultaneous contraction of inputs and expansion of outputs. Also, use of SBM instead of the more traditional CCR or BCC models allows the analysis to capture the nonradial reduction in inputs and nonradial increase in outputs; that is, the radial changes assumed in the CCR and BCC models may be inappropriate unless proportionality can be supported. For example, a bank may follow a policy of paying higher salaries to retain employees who excel in customer service, that is, incurring relatively more *noninterest expense*, but makes up for this by offering slightly lower rates on its deposit accounts (i.e., incurring relatively less *interest expense*). Similarly, the same bank may opt to focus its attention on the sale of those products and services that fetch more fee income than loans (i.e., earning relatively more *noninterest income*). In these examples, radial adjustments cannot adequately account for the bank's operational preferences. An equally important point to remember is that such operational preferences are likely to change over time in a dynamic world of business, thus making the choice of nonradial efficiency measures more appropriate for the real world. NSBM also has other desirable properties such as units invariance and strict monotonicity.

15.5.3.3 Data and Simulation

Empirical tests use a combination of actual data sourced from financial statements on the four core profit efficiency variables identified above for UAE domestic commercial banks, and simulated data on their key profit centers because such data are not available in the public domain. BankScope, the source for data on the core profit efficiency model, lists 18 commercial banks. After removing those banks with missing data, 15 domestic commercial banks comprise the sample in the fiscal year ending December 31, 2005 (based on the International Financial Reporting Standards), which was the most up-to-date reporting period found in BankScope at the time of writing this study.

To estimate profit center data that are not available to those outside the bank, the study assumes that the actual observed core bank-level expense and income data (as obtained from BankScope) are allocated among the three profit centers in randomly varying proportions. Thus, the data corresponding to the variables in Table 15.13 are simulated within predetermined parameters based on what can be observed. Details of this simulation process can be read in Avkiran (2009a).

In summary, the study first estimates the proportion of total outputs corresponding to each profit center by allowing the proportions to vary randomly in designated ranges. Then, the ratios that emerge from the actual observed data at the bank level are used to further disaggregate the simulated profit center data into interest income and noninterest income. A similar procedure is followed in the estimation of total inputs corresponding to each profit center by allowing the proportions to vary in the designated ranges, followed by disaggregating the simulated data to estimate profit center interest expenses and noninterest expenses. Intermediate inputs are allowed to vary in a predetermined range. On the other hand, intermediate outputs depend on observed bank-level net interest margins and simulated profit center intermediate inputs. The examples of intermediate products in this study, that is, referrals, are treated as discretionary variables where the profit center managers can control such activity. The results and analysis section demonstrates the differences between traditional DEA on the core profit efficiency model and the network model that accounts for sub-DMUs.

15.5.4 Results and Analysis

15.5.4.1 Profit Efficiency Using Traditional DEA (SBM)

Recapping, the core variables for each bank are *interest expense* (input), *noninterest expense* (input), *interest income* (output), and *noninterest income* (output). Initial calculations involve the profit efficiency of UAE domestic commercial banks using nonoriented, variable-returns-to-scale super SBM and observed data; in Table 15.14, this estimate is named the "bank black box score." Estimates on the sample of 15 banks show seven banks as efficient compared to the others in the sample. Using super-efficiency scores enables ranking among the efficient banks, where the National Bank of Umm Al-Qaiwain is leading the others by a considerable margin (see Chen 2004; Adler and Raveh 2008).

The right-hand side of Table 15.14 reports similar estimates but is based on simulated data, where profit centers are treated as independent sub-DMUs rather than interacting sub-DMUs. The idea of independent sub-DMUs was introduced earlier as part of the discussion on the separation model depicted in Fig. 15.2. A closer look reveals that the bank black box score is not correlated with the overall profit center score (Pearson's r equals -0.250 with a two-tailed significance level of 0.368). This implies that the weighted sum of the noninteracting divisional performances does not correspond to the black box score that represents the organization at aggregate level.

Table 15.14 Super-efficiency estimates for UAE banks and their profit centers (assuming independent centers)

Bank	Bank black box score	LAO score[a] (0.9156)	DCB score (0.0444)	MREL score (0.0400)	Overall profit center score[b]
National Bank of Umm Al-Qaiwain	1.6293	0.0104	3.8927	0.0383	0.1839
Bank of Sharjah	1.3844	1.3627	0.0263	15.7122	1.8774
National Bank of Abu Dhabi	1.3055	1.2665	2.7427	0.1666	1.2880
Mashreqbank	1.2209	0.2999	0.2028	1.1108	0.3280
First Gulf Bank	1.1654	1.4158	0.0669	0.2242	1.3082
Commercial Bank of Dubai	1.1162	0.1240	0.6181	0.1687	0.1478
Abu Dhabi Commercial Bank	1.0790	1.0793	0.3060	0.0700	1.0046
Union National Bank	0.9628	0.4076	0.0143	0.7169	0.4025
Emirates Bank International	0.8146	0.0025	0.5034	1.3516	0.0787
RAKBANK-National Bank of Ras Al-Khaimah	0.4599	0.0058	0.3061	0.1353	0.0243
National Bank of Dubai Public Joint Stock Company	0.3695	0.1544	1.3094	0.0462	0.2014
Commercial Bank International	0.3689	0.0137	0.2520	0.0770	0.0268
National Bank of Fujairah	0.3423	0.0012	0.0833	1.3518	0.0589
Arab Bank for Investment & Foreign Trade	0.2933	17.9763	0.0153	0.0178	16.4605
United Arab Bank	0.2514	0.0240	0.0666	0.0899	0.0286

LAO loans, advances, and overdrafts, *MREL* mortgaged real estate loans, *DCB* discounted commercial bills

[a] Using simulated data, these profit center scores represent the production processes depicted in Fig. 15.2 and divisional weights are shown in parentheses

[b] Overall profit center score is the weighted sum of divisional scores

15.5.4.2 Profit Efficiency Using Network SBM and Simulated Profit Center Data

Simulation is used to overcome lack of access to internal bank data needed to demonstrate NSBM. Nevertheless, the innovative approach to simulating data at the divisional level depends on actual observed data at the sector and individual bank levels. This brings a certain level of realism to the simulation exercise, although the current study is unable to draw conclusions on actual individual bank performances based on NSBM results. Table 15.15 shows the overall network efficiency estimates, a breakdown of profit center scores, and their corresponding projected inputs and outputs (see the footnote to the table).

Overall NSBM score is the weighted sum of divisional scores where the divisions are allowed to interact. Since none of the banks have all three divisions efficient, NSBM efficiency is not achieved (a similar relationship was observed by

Table 15.15 NSBM analysis of UAE banks (assuming interacting centers)

Bank	Overall NSBM score	Profit center scores[a]			Profit center projections measured as % change[b]					
					Inputs			Outputs		
		LAO (0.9156)	DCB (0.0444)	MREL (0.0400)	Ie	Nie	Nie*	Ii	Nii	Nii*
National Bank of Abu Dhabi	0.964	1.000	1.000	0.332	−69	−51	0	0	43	n/a
First Gulf Bank	0.915	1.000	0.273	1.000	−26	0	n/a	98	339	n/a
Bank of Sharjah	0.788	1.000	0.106	1.000	0	−41	n/a	1,212	84	n/a
Abu Dhabi Commercial Bank	0.731	1.000	0.452	0.111	0	−36	n/a	127	37	n/a
					−18	0	0	559	884	n/a
Arab Bank for Investment & Foreign Trade	0.558	1.000	0.040	1.000	0	−37	n/a	993	2,907	n/a
Union National Bank	0.516	0.608	0.142	0.678	−33	−37	n/a	0	14	−12
					0	−15	n/a	728	380	n/a
					−53	−7	−12	7	0	n/a
Mashreqbank	0.289	0.270	0.203	1.000	−39	−76	n/a	114	2	0
					−67	−85	n/a	36	0	n/a
National Bank of Dubai Public Joint Stock Company	0.102	0.093	1.000	0.058	−69	−72	n/a	0	441	−83
					−82	−66	−83	94	614	n/a
Commercial Bank of Dubai	0.085	0.036	0.618	0.563	−95	−97	n/a	29	0	−25
					−4	−72	n/a	0	0	n/a
					−2	−40	−24	0	81	n/a
National Bank of Umm Al-Qaiwain	0.059	0.055	1.000	0.054	−28	−61	n/a	548	1,262	−99
					−87	−93	−99	22	157	n/a
Commercial Bank International	0.054	0.050	0.252	0.104	−75	−82	n/a	110	551	−96
					−65	−53	n/a	30	94	n/a
					−82	−86	−96	0	109	n/a
RAKBANK-National Bank of Ras Al-Khaimah	0.047	0.041	0.306	0.160	−73	−90	n/a	41	655	−81
					−43	−85	n/a	0	35	n/a
					−54	−82	−80	0	199	n/a
United Arab Bank	0.043	0.040	0.067	0.115	−84	−77	n/a	8	750	−87
					−88	−68	n/a	39	434	n/a
					−75	−63	−87	0	341	n/a
Emirates Bank International	0.008	0.007	0.503	1.000	−54	−81	n/a	5,093	4,430	0
					−32	−67	n/a	0	0	n/a
National Bank of Fujairah	0.006	0.005	0.083	0.718	−75	−79	n/a	2,049	7,321	145
					−79	−67	n/a	128	311	n/a
					−7	−18	145	0	43	n/a

Ie interest expense, *Nie* noninterest expense, *Nie** noninterest expense as intermediate input to MREL, *Ii* interest income, *Nii* noninterest income, *Nii** noninterest income as intermediate output from LAO, *LAO* loans, advances, and overdrafts, *MREL* mortgaged real estate loans, *DCB* discounted commercial bills

[a] These profit center scores represent the production processes depicted in Fig. 15.1 involving interacting centers (weights are shown in parentheses)

[b] Profit center projections (rounded to the nearest integer) are reported in the order they appear. For example, for Abu Dhabi Commercial Bank, DCB projections would be followed by MREL projections. However, since DCB does not have any intermediate products, n/a is inserted for Nie* and Nii*. For the same bank, the second row of projections indicate n/a for Nii* because MREL is not linked to an intermediate output (however, it is linked to an intermediate input with a projected change of zero)

Table 15.16 Comparison of rank correlations across SBM, NSBM, and multiple simulations

	NSBM		
Kendal's tau-b	Simulation 1	Simulation 2	Simulation 3
Simulation 1		0.295(0.125)	0.000(1.000)
Simulation 2			0.077(0.692)
SBM	0.390(0.042)	0.371(0.054)	0.230(0.234)

Lewis and Sexton 2004). Slacks for divisional level inputs and outputs indicate some of the hidden inefficiencies that are not normally visible with traditional DEA. For example, the Union National Bank, whose three profit centers are inefficient, would particularly benefit by focusing its attention on the DCB profit center. The DCB profit center can increase its interest income and noninterest income by 728% and 380% respectively, while reducing noninterest expense by 15%. Exposing this kind of detail at the divisional level is the main use of nonoriented NSBM that can help in targeting remedial action.

Further insight into divisional inefficiencies can be gained by focusing on intermediate products. The column headed Nii* indicates seven banks (which are in bold font in Table 15.15) whose LAO profit centers would be better off reducing intermediate outputs, that is, noninterest income from referrals, with corresponding similar reductions in the intermediate inputs of noninterest expense for MREL profit centers. An exception to this pattern is the National Bank of Fujairah where the projections suggest an increase of 145% in intermediate products. The ostensibly proportional projections for intermediate products in this case are explained by the specially formulated relationship between intermediate inputs and intermediate outputs.

An equally important comparison is the rank correlation between the black box scores in Table 15.14 and the overall NSBM scores in Table 15.15. Kendal's tau-b of 0.390(0.042) highlights the substantial differences between traditional SBM and NSBM (see Table 15.16). Table 15.16 also depicts the robustness of the simulation method. Starting from the premise that the original simulation generated random numbers, the simulation is repeated two more times. Insignificant rank correlations between SBM and NSBM scores, as well as insignificant rank correlations among NSBM scores from different simulations indicate a successful attempt at generating independent samples.

15.5.5 Concluding Remarks

The key motivation for this study is that traditional DEA does not provide adequate detail for management to identify the specific sources of inefficiency embedded in interactions among divisions of an organization. In an age of global economy characterized by cross-border, fierce competition, business organizations are increasingly forced to look for new ways of identifying inefficiencies in order to

sustain their competitive advantage in the market. The study illustrates how NSBM can address the above need by providing insight to the specific sources of organizational process inefficiency in the context of UAE domestic commercial banks. By helping management open the "black box" of production, NSBM gives access to the underlying diagnostic information in divisions (profit centers) that would otherwise remain undiscovered.

The new approach developed and applied in this study has global appeal. It can be used in those countries where an industry is known to suffer from relatively high levels of inefficiency, as well as in countries where the overall measured inefficiencies are low but more focused strategies are needed to identify and eliminate the remaining inefficiencies. Developing reliable performance measures would also assist in managerial decision-making involving mergers where executives typically search for synergies that are often enjoyed by acquiring divisions that provide complementary services.

15.6 Epilogue

Productivity analysis in the service sector has lagged behind the manufacturing sector for many decades. This has been partly due to the emergence of management science from the manufacturing domain following the Industrial Revolution. In the first half of the twentieth century, economists were pre-occupied with the manufacturing sector. It was not until the second half of the twentieth century that economists began to acknowledge the importance of the service sector. The twenty-first century is shaping as the coming of age of productivity analysis in the service sector.

Productivity analysis in the service sector originally borrowed from the manufacturing sector with less than satisfactory results. Attempts to develop productivity measures specific to the service sector have often run into difficulty due to the intangible nature of service. The productivity ratios devised fell short of accounting for the complex relationships between inputs and outputs. In this chapter, it has been demonstrated in various case studies that it is possible to capture the interplay between multiple inputs and multiple outputs through DEA.

We ought to bear in mind that the primary purpose of productivity analysis is *improvement*. That is, the process of measurement has limited intrinsic value if the findings are not used to reduce inefficiencies – an important process for profitability as well as for running globally sustainable economies. For best results, DEA should be interpreted in the context of other measurement systems in place, such as balanced scorecard. In fact, the most astute application of DEA is to use it as a starting point for more in-depth investigation before attempting to raise productivity.

In theory, productivity can be improved in five different ways (Sandler 1982, p. 153):

(a) Increase output more than input
(b) Increase output without increasing input

(c) Decrease output but decrease input more
(d) Maintain same output but decrease input
(e) Increase output and decrease input

Regardless of the preferred path to productivity improvement, it is first necessary to ensure that the organizational commitment is in place. This translates into awareness by management that productivity improvement is not simply dependent on performance of line staff but on all levels of management. In order to secure commitment, it makes sense to highlight the benefits of productivity improvement (and its measures) to the organization and the individuals who work there. In this respect, it is important to devise comprehensible and actionable measures that can be demonstrated to tie in with the selected improvement goals.

We end this chapter by going over the main pitfalls to avoid in application of DEA (see also chapter 11 in Avkiran 2006 and Dyson et al. 2001), and an application checklist:

1. *Placing nonhomogeneous DMUs in the same sample*. This can distort results as DEA assumes DMUs to be homogeneous. That is, DMUs are assumed to be involved in similar activities, produce a common set of outputs, and operate in comparable environments. Where large numbers of DMUs are available, clustering of similar DMUs and running separate DEA can alleviate this problem.
2. *Assuming all inputs and outputs are within managerial discretion*. Certain exogenous factors such as environmental conditions and regulatory framework are outside the managerial sphere of influence. There may also be constraining factors such as market demand that is partially controllable by management. Two potential modeling approaches are either to include the exogenous factors as inputs and run an output maximizing DEA, or to include exogenous factors as outputs and run an input minimizing DEA. Significant differences in the operating environment of DMUs can be captured by employing multiple-stage analyses that include regression.
3. *Assuming inputs are bad and outputs are good for your efficiency score*. DEA assumes data to be isotonic, that is, efficiency drops as inputs rise, and efficiency rises as outputs increase. Examples of anti-isotonic variables include pollution (an undesirable output) and number of competitors in the catchment area of a business (an inhibiting input).
4. *Using a small sample size in relation to variables*. As the number of variables in DEA rise, level of discrimination among DMUs falls. The minimum sample size is often calculated as twice the product of number of inputs and number of outputs, or three times the sum of number of inputs and number of outputs. A third rule-of-thumb suggests that there is too much loss of discriminatory power when more than one-third of the DMUs emerge as efficient.

Finally, an application checklist is presented for the benefit of those who are keen to implement DEA:

1. Define the DMU to be studied. How many of these units are there?
2. Identify the business drivers (outputs) critical to success of the DMU.

3. Identify the key resources (inputs) that support the key business drivers. Here, a process analysis can provide direction.
4. Are data on the key outputs/inputs collected in a regular and consistent manner?
5. Is there a particular managerial perspective that you would like to bring to the analysis? For example, service volume, service quality, profit efficiency, overall efficiency, and so on.
6. Are you interested in output maximization or input minimization as the main objective? Or, would a nonoriented approach serve better your managerial perspective?
7. Is there any evidence of variable-returns-to-scale in the units to be analyzed?
8. Run the DEA calculations and determine the units reported as inefficient.
9. Are the inefficient units consistently inefficient over time?
10. Are inefficient units measured as efficient when analyzed under different methods? If so, determine why. For example, see whether environmental factors have been adequately considered.
11. Identify the best practice unit in the sample.
12. Identify the potential improvements for the inefficient units, as well as their corresponding reference sets that will give direction to the new process of emulation.
13. Are there organizational constraints to implementation of the potential improvements? To answer this, revisit outputs and inputs studied and relate them to existing budgets and stakeholders.
14. Communicate the results of the first round of DEA to those mangers that will be affected by the projected changes to inputs and outputs. Invite feedback on the observed differences in the performance of different DMUs.

If the previous step uncovers significant variables that were omitted, a second round of DEA should be undertaken after such variables are satisfactorily incorporated into the productivity model.

References

Adler N, Raveh A. Presenting DEA graphically. OMEGA The Int J Manage Sci. 2008;36:715–29.

Ahn T, Arnold V, Charnes A, Cooper WW. DEA and ratio efficiency analyses for public institutions of higher learning in Texas. Res Govt Nonprofit Accounting. 1989;5:165–85.

Ahn T, Charnes A, Cooper WW. Some statistical and DEA evaluations of relative efficiencies of public and private institutions of higher learning. Socioecon Plann Sci. 1988;22(6):259–69.

Ahn T, Seiford LM. Sensitivity of DEA models and variable sets in a hypothesis test setting: the efficiency of university operations. In: Ijiri Y, editor. Creative and innovative approaches to the science of management. Wesport: Quorum Books; 1993. p. 191–208.

Ahn T. Efficiency and related issues in higher education: a data envelopment analysis approach. Doctoral dissertation. The University of Texas at Austin; 1987.

Amado CAF, Dyson RG. Exploring the use of DEA for formative evaluation in primary diabetes care: an application to compare English practices. J Oper Res Soc. 2009;60(11):1469–82.

Anderson L, Walberg HJ. Data envelopment analysis. In: Keeves JP, editor. Educational research, methodology, and measurement: an international handbook. Adelaide: Flinders University of South Australia; 1997. p. 498–503.
Andrews L, Aungles P, Baker S, Sarris A. Characteristics and performance of higher education institutions (a preliminary investigation). Canberra: Department of Employment, Education, Training and Youth Affairs; 1998.
Avkiran NK. Productivity analysis in the service sector with data envelopment analysis. Brisbane: Avkiran NK; 2006.
Avkiran NK. Investigating technical and scale efficiencies of Australian universities through data envelopment analysis. Socioecon Plann Sci. 2001;35(1):57–80.
Avkiran NK. Monitoring hotel performance. J Asia Pac Business. 2002;4(1):51–66.
Avkiran NK. Opening the black Box of efficiency analysis: an illustration with UAE banks. OMEGA The Int J Manage Sci. 2009a;37(4):930–41.
Avkiran NK. Removing the impact of environment with units-invariant efficient frontier analysis: an illustrative case study with intertemporal panel data. OMEGA The Int J Manage Sci. 2009b;37(3):535–44.
Avkiran NK, Thoraneenitiyan N. Purging data before productivity analysis. J Bus Res. 2010;63 (3):294–302.
Avkiran NK. The evidence on efficiency gains: the role of mergers and the benefits to the public. J Bank Finance. 1999;23:991–1013.
Avkiran NK. Rising productivity of Australian trading banks under deregulation 1986–1995. J Econ Finance. 2000;24:122–40.
Ball SD, Johnson K, Slattery P. Labour productivity in hotels: an empirical analysis. Int J Hospitality Manage. 1986;5(3):141–7.
Banker RD, Thrall RM. Estimation of returns to scale using data envelopment analysis. Eur J Oper Res. 1992;62:74–84.
Banker RD. Estimating most productive scale size using data envelopment analysis. Eur J Oper Res. 1984;17:35–44.
Barrow M, Wagstaff A. Efficiency measurement in the public sector: an appraisal. Fisc Stud. 1989;10(1):72–97.
Beasley JE. Comparing university departments. OMEGA The Int J Manage Sci. 1990;18 (2):171–83.
Bhattacharyya A, Lovell CAK, Sahay P. The impact of liberalization on the productive efficiency of Indian commercial banks. Eur J Oper Res. 1997;98:332–45.
Birch S, Maynard A. Performance indicators and performance assessment in the UK national health service: implications for management and planning. Int J Health Manage. 1986;1:143–56.
Breu TM, Raab RL. Efficiency and Perceived Quality of the Nation's top 25 National Universities and National Liberal Arts Colleges: An Application of Data Envelopment Analysis to Higher Education. Socioecon Plann Sci. 1994;28(1):33–45.
Brockett PL, Charnes A, Cooper WW, Huang ZM, Sun DB. Data transformations in DEA cone ratio envelopment approaches for monitoring bank performances. Eur J Oper Res. 1997;98:250–68.
Business Queensland. Brisbane: Business Queensland Book of Lists; 1997.
Butterfield RW. A quality strategy for service organizations. Qual Prog. 1987;20(12):40–2.
Cave M, Hanney S, Kogan M. The Use of performance indicators in higher education: a critical analysis of developing practice. London: Jessica Kingsley Publishers; 1991.
Central Bank of UAE. Annual Report; 2006. http://www.centralbank.ae/en/index.php?option=com_content&view=article&id=100&Itemid=88.
Charnes A, Cooper WW, Huang ZM, Sun DB. Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks. J Econom. 1990;46:73–91.
Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res. 1978;2:429–44.

Chen Y. Ranking efficient units in DEA. OMEGA The Int J Manage Sci. 2004;32:213–9.
Coelli T, Rao DSP, Battese GE. An introduction to efficiency and productivity analysis. Boston: Kluwer Academic Publishers; 1998.
DEETYA. The composite index: allocation of the research quantum to Australian universities. Canberra: Research Branch, Higher Education Division; 1997.
Denton GA, White B. Implementing a balanced-scorecard approach to managing hotel operations. Cornell Hotel Restaur Adm Q. 2000;41(1):94–107.
Dyson RG, Allen R, Camanho AS, Podinovski VV, Sarrico CS, Shale EA. Pitfalls and protocols in DEA. Eur J Oper Res. 2001;132:245–59.
Dyson RG, Thanassoulis E. Reducing weight flexibility in data envelopment analysis. J Oper Res Soc. 1988;39(6):563–76.
Färe R, Grosskopf S, Whittaker G. Network DEA. In: Zhu J, Cook WD, editors. Modeling data irregularities and structural complexities in data envelopment analysis. New York: Springer Science Business Media; 2007.
Färe R, Grosskopf S. Intertemporal production frontiers: with dynamic DEA. Boston: Kluwer Academic Publishers; 1996.
Färe R, Grosskopf S. Network DEA. Socioecon Plann Sci. 2000;34:35–49.
Flynn N. Performance measurement in public sector services. Policy Polit. 1986;14(3):389–404.
Geller AN. Tracking the critical success factors for hotel companies. Cornell Hotel Restaur Adm Q. 1985;25(4):76–81.
Gillen D, Lall A. Developing measures of airport productivity and performance: an application of data envelopment analysis. Transp Res E. 1997;33(4):261–73.
Guerrier Y, Lockwood AJ. Work flexibility in hotels. In: Johnston R, editors. Proceedings of the Annual International Conference of the Operations Management Association. Warwick: University of Warwick; 1988. Pp. 351–365.
Haywood KM. Assessing the quality of hospitality services. Int J Hospitality Manage. 1983;2 (4):165–77.
Johnes G. Research performance indications in the university sector. High Edu Quart. 1988;42 (1):54–71.
Johnes J. Performance assessment in higher education in Britain. Eur J Oper Res. 1996;89:18–33.
Johns N, Wheeler K. Productivity and performance measurement and monitoring. In: Teare R, Boer A, editors. Strategic hospitality management. London: Cassell; 1991. p. 45–71.
Jones P. Quality, capacity and productivity in service industries. Int J Hospitality Manage. 1988;7 (2):104–12.
Leightner JE, Lovell CAK. The impact of financial liberalization on the performance of Thai banks. J Econ Bus. 1998;50:115–31.
Lewis HF, Sexton TR. Network DEA: efficiency analysis of organizations with complex internal structure. Comput Oper Res. 2004;31:1365–410.
Lindsay AW. Institutional performance in higher education: the efficiency dimension. Rev Educ Res. 1982;52(2):175–99.
Mclaughlin CP, Coffey S. Measuring productivity in services. In: Lovelock CH, editor. Managing services: marketing, operations, and human resources. 2nd ed. New Jersey: Prentice-Hall; 1992. p. 103–17.
Messenger SJ, Mugomeza C. An exploratory study of productivity and performance measurement in Zimbabwean hotels. Int J Contemp Hospit Manag. 1995;7(5):V–VII.
Mill RC. Productivity in service organisations. In: Jones P, editor. Management in service industries. London: Pitman; 1989. p. 275–88.
Miller SM, Noulas AG. The technical efficiency of large bank production. J Bank Finance. 1996;20:495–509.
Morita H, Hirokawa K, Zhu J. A slack-based measure of efficiency in context-dependent data envelopment analysis. OMEGA The Int J Manage Sci. 2005;33:357–62.
Nunamaker TR. Using data envelopment analysis to measure the efficiency of Non-profit organizations: a critical evaluation. Managerial Decis Econom. 1985;6(1):50–8.

Olsen M, Crawford-Welch S, Tse E. The Global hospitality industry of the 1990s. In: Teare R, Boer A, editors. Strategic hospitality management. London: Cassell; 1991. p. 213–26.
REIQ (Real Estate Institute of Queensland). Queensland residential real estate agency practice profitability report 1992/93; 1994.
Renaghan LM. A new marketing Mix for the hospitality industry. Cornell Hotel Restaur Adm Q. 1981;22(2):31–5.
Sandler M. Productivity Measurement and Improvement in the Hospitality Industry. In: Pizam A et al., editors. The practice of hospitality management. Connecticut: AVI Publishing Co.; 1982. p. 153–64.
Schroeder RG. Operations management: decision making in the operations function. 2nd ed. New York: McGraw-Hill; 1985.
Sealey Jr CW, Lindley JT. Inputs, outputs, and a theory of production and cost at depository financial institutions. J Finance. 1977;32:1251–66.
Sharma KR, Leung P, Chen HL. Economic efficiency and optimum stocking densities in fish polyculture: an application of data envelopment analysis (DEA) to Chinese fish farms. Aquaculture. 1999;180(3–4):207–21.
Sherman HD, Zhu J. Benchmarking with quality-adjusted DEA (Q-DEA) to seek lower-cost high-quality service: evidence from a U.S. Bank application. Ann Oper Res. 2006;145:301–19.
Sherman HD. Service organization productivity management. Hamilton: The Society of Management Accountants of Canada; 1988.
Solomons D. Divisional performance: measurement and control. Homewood: R D Irwin; 1968.
Sturm JE, Williams B. Foreign banks entry deregulation and bank efficiency: lessons from the Australian experience. J Bank Finance. 2004;28:1775–99.
Thomas JA. The productive school: a system analysis approach to educational administration. New York: Wiley; 1974.
Tomkins C, Green R. An experiment in the use of data envelopment analysis for evaluating the efficiency of UK university departments of accounting. Financ Accountability and Manage. 1988;4(2):147–64.
tone K, Tsutsui M. Network DEA: a slacks-based measure approach. Eur J Oper Res. 2009;197:243–52.
Tone K. A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res. 2001;130:498–509.
Vitikainen K, Street A, Linna M. Estimation of hospital efficiency – do different definitions and casemix measures for hospital output affect the results? Health Policy. 2009;89(2):149–59.
Witt CA, Witt SF. Why productivity in the hotel sector is low. J Contem Hospitality Manage. 1989;1(2):28–34.
Witt SF, Moutinho L. Tourism marketing and management handbook. Hertfordshire: Prentice Hall; 1994.
Wong YHB, Beasley JE. Restricting weight flexibility in data envelopment analysis. J Oper Res Soc. 1990;41(9):829–35.
Yeh QJ. The application of data envelopment analysis in conjunction with financial ratios for bank performance evaluation. J Oper Res Soc. 1996;47:980–8.
Yue P. Data envelopment analysis and commercial bank performance: a primer with applications to Missouri banks. Federal Reserve Bank of St Louis Rev. 1992;74:31–45.

Chapter 16
Health-Care Applications: From Hospitals to Physicians, from Productive Efficiency to Quality Frontiers

Jon A. Chilingerian and H. David Sherman

Abstract This chapter focuses on health-care applications of DEA. The paper begins with a brief history of health applications and discusses some of the models and the motivation behind the applications. Using DEA to develop quality frontiers in health services is offered as a new and promising direction. The paper concludes with an eight-step application procedure and list of do's and don'ts when applying DEA to health services.

Keywords Health Services Research • Physicians • Hospitals • HMOs • Frontier Analysis • Health-Care Management • DEA • Performance • Efficiency • Quality

16.1 Introduction

Throughout the world, health-care delivery systems have been under increasing pressure to improve performance: that is, to control health-care costs while guaranteeing high-quality services and better access to care. Improvements in health-care performance are important because they can boost the well-being, as well as the standard of living and the economic growth of any nation. The quest for high performance in health care has been a difficult and intractable problem historically. Efforts to reduce costs and improve service quality and access have been only marginally successful (Georgopoulos 1986; Newhouse 1994; Shortell and Kaluzny 2000).

Although no theoretically correct or precise measures of "health" exist, there is a great deal of interest in studying and understanding health-care costs, outcomes, and utilization. This interest in understanding health outcomes is associated with our continual desire to improve health care. While we may have no precise measures of

H.D. Sherman (✉)
College of Business Administration, Northeastern University, Boston, MA 02115, USA
e-mail: H.Sherman@NEU.edu

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8_16,

health-care performance, the desire to improve health care can be seen in the unrelenting increase in the quantity of health products and services available to patients. Yet, even with these increases, health care seems to offer fewer perceived benefits to patients in relation to their perceived sacrifices (see Anderson et al. 2003; Bristol Royal Infirmary Inquiry Final Report 2002; Newhouse 2002).

As one economist has proclaimed, "Despite the lack of a summary measure of its efficiency, many seem convinced that the industry's performance falls short" (Newhouse 2002: 14). For example, one medical center in England received special funding to become a Supra Regional Service center for pediatric cardiac surgical care (Bristol Royal Infirmary Inquiry Final Report 2002). This program received funding over a 14-year period, despite significantly higher mortality and morbidity rates and poor physician performance. Should a clinical manager have detected these poor practices sooner? In addition, what internal control systems are available to measure and evaluate individual physician performance?

In 2000, although the USA spent far more on health care per capita than any other country in the world, the number of physicians per 1,000 population, primary care visits per capita, acute beds per capita, hospital admissions per capita, and hospital days per capita were below the median of most other developed countries (Anderson et al. 2003). Are quality and productive efficiency of some national health-care systems really lower than that of many other industrialized nations? What accounts for performance differences?

Twenty years ago, research studies that questioned the patterns or cost of care rarely attempted to estimate the amount and sources of inefficiency and poor performance. Today that has changed. Several hundred productive efficiency studies have been conducted in countries such as Austria, Finland, the Netherlands, Norway, Spain, Sweden, the UK, and the USA. These studies have found evidence that technical inefficiency in these systems is significant. For example, in the USA, such inefficiencies result in billions of "wasted" US dollars (USD) each year. While other statistical techniques have also been used, Data Envelopment Analysis (DEA) has become the researchers' method of choice for finding best practices and evaluating productive inefficiency.

While there are several techniques for estimating best practices such as stochastic frontier analysis, fixed effects regression models, and simple ratios, DEA has become a preferred methodology when evaluating health-care providers. DEA is a methodology that estimates the degree to which observed performance reaches its potential and/or indicates how well resources have been utilized (see Chap. 1 in this handbook). Benchmarked against actual behavior of decision-making units (DMUs), DEA finds a best practices frontier – i.e., the best attainable results observed in practice – rather than central tendencies. The distance of the DMUs to the frontier provides a measure of overall performance. Although criticized by some health economists (see Newhouse 1994), the late Harvey Leibenstein praised DEA as a primary method for measuring and partitioning X-inefficiency (Leibenstein and Maital 1992).

DEA offers many advantages when applied to the problem of evaluating the performance of health-care organizations. First, the models are nonparametric

and do not require a functional form to be prescribed explicitly, i.e., linear, nonlinear, log-linear, and so on (Charnes et al. 1994). Second, unlike statistical regressions that average performance across many service providers, DEA estimates best practice by evaluating the performance behavior of each individual provider, comparing each provider with every other provider in the sample. The analysis identifies the amount of the performance deficiency and the source. Third, unlike regression and other statistical methods, DEA can handle multiple variables, so the analysis produces a single, overall measure of the best results observed.

Finally, to identify those providers who achieved the best results, DEA groups providers into homogeneous subgroups. Providers that lie on the frontier achieved the best possible results and are rated 100% efficient. Providers that do not lie on the surface underperformed, and their performance is measured by their distance from the frontier. The analysis not only provides a measure of their relative performance but also uncovers subgroups of providers similar in their behavior or similar in the focus of their attention to performance.

This chapter focuses on health-care applications of DEA. We begin with a brief history of health-care applications. Next, several health-care models and approaches are discussed, followed by new directions for future studies. We conclude with a summary of do's and don'ts when applying DEA to evaluate the performance of health-care organizations.

16.2 Brief Background and History

For health-care organizations, finding a single overall measure of performance has been difficult. Goals of most health-care services are multiple, conflicting, intangible, vague, and complex. Virtually, every study of efficiency could be criticized for failing to look at quality, clinical innovation, or the changing nature of the services (Newhouse 1994). The worldwide demand to provide health-care services at lower cost made this an ideal focus for DEA research, and health care (like banking) has had innumerable DEA applications.

The first application of DEA in health care began with H. David Sherman's Doctoral dissertation in 1981. W.W. "Bill" Cooper had been a professor at Harvard Business School in 1979, David Sherman was a doctoral candidate and he and Rajiv Banker happened to take Bill Cooper's seminar. Bill Cooper mentioned DEA as a new technique, and both Banker and Sherman used that technique in their dissertations. Working with the Massachusetts Rate Setting Commission to ensure relevance in the complex task of evaluating hospital performance, David Sherman applied DEA to evaluate the performance of medical and surgical departments in 15 hospitals. When Dr. Sherman became a professor at MIT in 1980, his doctoral student and research assistant, Jon Chilingerian helped him to compare DEA results with other statistical models and to look for interesting and novel health-care applications.

In 1983, Nunamaker published the first health application using DEA to study nursing services (Nunamaker 1983). In 1984, the second DEA paper (Sherman 1984) evaluated the medical and surgical departments of seven hospitals. By 1997, Hollingsworth et al. (1999) counted 91 DEA studies in health care. The health applications include the following: health districts, HMOs, mental health programs, hospitals, nursing homes, acute physicians, and primary care physicians.

DEA applications in health have been evolving over the last two decades. As access to bases and information technology have improved, so have the quality of the studies. In the next section, we review some of the DEA literature in health care starting with acute hospitals, nursing homes, other health organizations, and physicians.

16.2.1 Acute General Hospitals and Academic Medical Centers

Acute hospitals have received the most research attention using DEA. These hospital studies use DEA alone. They measure overall technical efficiency using the CCR model. Finally, they define outputs as patient days or discharges and do not measure clinical outcomes. A comprehensive review of efficiency studies in health care by Hollingsworth et al. (1999) found systematic differences among the average efficiency and range of DEA scores by ownership type and national hospital systems. For example, public hospitals had the highest mean efficiency scores (0.96) and not-for-profit hospitals had lower mean DEA scores (0.80). When comparing US studies with studies from other European countries, Hollingsworth et al. found a greater potential for improvement in the USA with an average efficiency score of 0.85, and a range of 0.60–0.98, in contrast to Europe with an average efficiency score of 0.91, and a range of 0.88–0.93.

Although comparing DEA scores among various hospital studies is useful for hypothesis generation, there are many limitations. First, these studies used different input and output measures during different time periods. Second, the distribution of DEA scores is so skewed, (given the huge spike of efficient units), that reliance on the usual measures of central tendency will be misleading. By excluding the efficient units, the average inefficiency score may be a more reasonable comparison. Third, the output measures in these studies are vastly different. Many did not use the same type of case-mix-adjusted data, and some studies used crude case-mix-adjusted data such as age-adjusted discharges, so the results are likely affected by unaccounted case-mix differences.

One recent study of 22 hospitals in the National Health Service in the UK used a four output, five input model (Kerr et al. 1999). The outputs were defined as the following: (1) Surgical inpatients and visits, (2) Medical inpatients and visits, Obstetrics/Gynecology patients and visits, Accidents and Emergency visits. Without knowing the complexity and severity of patients, raw measures of output will lead to distorted results. If Hospital A receives a lower DEA score because Hospital A admits more “fevers of unknown origin,” and performs more

combined liver–kidney transplants, hip replacements, and coronary bypass grafts and Hospital B has more tooth extractions, vaginal deliveries without complications, and circumcisions, it is an unfair comparison.

Acute hospitals are among the most complex organizations to manage. Periodically, there are random fluctuations and chaos is everywhere – the emergency room, the operating rooms, the intensive care units, and the like. Hospital studies are complex because of the amount of input and output information needed to describe the clinical activities and services and the patients' trajectories. Most of care programs are not really under the control of the hospital manager (see Chilingerian and Glavin 1994; Chilingerian and Sherman 1990). Therefore, traditional DEA hospital studies may not have been useful to practicing managers.

Most of the hospital studies have merely illustrated DEA as a methodology and demonstrated its potential. Unfortunately, they use very different hospital production models. Some combine patient days with discharges as outputs, and others separate the manager-controlled production process from the clinical-controlled process. Few studies have tested any clinical or organizational theory.

Evaluating acute hospitals requires a large and complex DEA model. To identify underperforming hospitals DEA makes nontestable assumptions such as no random fluctuations, no measurement errors, no omitted outputs and output homogeneity (see Newhouse 1994). If additional inputs would improve quality, omitting an output variable such as the quantity of case-mix-adjusted mortalities can distort DEA results. Perhaps bringing DEA inside the hospital to compare departments, care programs, care teams, diseases, specific procedures and physicians' practice patterns would be more useful for practice. One might conclude from this that hospital-level comparisons are not the best application for DEA.

Nevertheless, there have been innovative hospital-level studies potentially useful for policy makers. For example, dozens of DEA papers have focused on the association between hospital ownership and technical inefficiency studying several thousand hospitals as DMUs (see for example, Burgess and Wilson 1996). DEA studies have also focused on critical health policy issues such as the following: regional variations (Perez 1992), rural hospital closures (Ozcan and Lynch 1992), urban hospital closures (Lynch and Ozcan 1994), hospital consolidations (Luke and Ozcan 1995), and rural hospital performance (Ferrier Ferrier and Valdanis 1996). O'Neill (1998) has recently made an important methodological contribution by developing an interesting DEA performance measure that allows policy makers to make fair comparisons of teaching and nonteaching hospitals. The DEA work on hospital performance continues to be important and needs more development.

16.2.2 Nursing Homes

Sexton et al. and Nyman and Bricker published the first two DEA studies of nursing homes in 1989. Sexton et al. ran a model that relied on two output measures (Medicaid and Other) and six inputs that only included labor; consequently,

the study had some limitations. Nyman and Bricker (1989) used DEA to study 195 for profit (FP) and not-for-profit (NFP) US nursing homes. Employing four categories of labor hours as inputs (e.g., nursing hours, social service hours, therapist hours, and other hours), and five outputs (skilled nursing patients (SNF), intermediate care patients (ICF), personal care patients, residential care patients, and limited care patients), they regressed the DEA scores in an ordinary least squares regression analysis and reported that NFP nursing homes were more efficient.

Another study in the USA by Nyman et al. (1990) investigated the technical efficiency of 296 nursing homes producing only intermediate care, no skilled nursing care, and relying on 11 labor inputs. They defined only one output, the quantity of intermediate care patients produced. Fizel and Nunnikhoven (1993) investigated the efficiency of US nursing home chains ignoring nonlabor inputs and focusing on two outputs: the quantity of intermediate and skilled nursing patients. They study found that chains were more efficient. Kooreman's 1994 study of 292 nursing homes in the Netherlands utilized six labor inputs and four case-mix-adjusted outputs: the quantity of physically disabled patients, quantity of psychogeriatrically disabled, quantity of full care, and quantity of day care patients. He found that 50% were operating efficiently, and the inefficient homes used 13% more labor inputs per unit of output, and quality and efficiency seemed to be going in the opposite direction.

A study of 461 nursing homes by Rosko et al., in 1995 employed five labor inputs, and two outputs, ICF patients and SNF patients. The study found that the variables associated with nursing home efficiency were managerial and environmental. Differences in efficiency were not associated with quality measures.

The nursing home studies are among the better applications of DEA. Although they encounter the same case-mix problems when modeling outputs, most of these studies regress the DEA scores to identify the variables associated with inefficiency. Since the outputs are often adjusted by payment types, or are crude patient types, many of these studies controlled for quality, patient characteristics, while exploring effects of ownership, operating environment and strategic choice on performance. These studies have found that managerial and environmental variables (the location of the home, nurse training, size of the homes, and wage rates) are strongly associated with the DEA scores, rather than quality of care, or patient mix.

16.2.3 Department Level, Team-Level, and General Health-Care Studies

In addition to acute care hospitals and nursing homes, DEA has been applied in a wide variety of health services and activities. For example, previous work has studied the productive efficiency of the following: Obstetrics units (Finkler and Wirtschafter 1993), pharmacies (Fare et al. 1994), intensive care units

(Puig-Junoy 1998), organ procurement programs (Ozcan et al. 1999), and dialysis centers (Ozgen and Ozcan 2002). All of these studies have used DEA to measure technical efficiency using basic models.

Ozcan et al.'s exploratory study conducted in 1999 on 64 organ procurement organizations employed four inputs (a capital proxy, FTE development labor, FTE Other Labor, and operating expenses), and two outputs (kidneys recovered, extra-renal organs recovered). The study found some evidence of scale efficiency not only did the larger programs produce 2.5 times more outputs, the average DEA efficiency scores were 95% for the large programs versus 79% for the smaller programs.

In 2000, a paper comparing the efficiency of 585 HMOs from 1985 to 1994 computed the estimates using data envelopment analysis, stochastic frontier analysis, and fixed effects regression modeling (Bryce et al. 2000). Unfortunately, they relied on a single output (total member-years of coverage) four input (hospital days, ambulatory visits, administrative expenses, and other expenses) model. They concluded that the three techniques identify different firms as more efficient.

In 2002, Ozgen and Ozcan reported a study of 791 dialysis centers in the USA. Constructing a multiple output (quantity of dialysis treatments, dialysis training, and home visits) and multiple input model (including clinical providers by type, other staff, operating expenses, and the number of dialysis machines – a proxy for capital), the study evaluated pure technical efficiency assuming variable returns to scale. The study found that the wide variations in efficiency were associated with ownership status – for-profit dialysis centers were less inefficient than not-for-profit ones.

16.2.4 Physician-Level Studies

A new and interesting development in health care has been taking DEA down to the workshop-level, focusing on individual physician performance. The first application of DEA at the individual physician level was in 1989; the analysis identified the nontechnical factors associated with technical efficiency of surgeons and internists (Chilingerian 1989).

Physician studies have developed new conceptual models of clinical efficiency (Chilingerian and Sherman 1990), as well as using DEA to explore new areas such as the following: most productive scale size of physician panels (Chilingerian 1995); benchmarking primary care gatekeepers (Chilingerian and Sherman 1996); and preferred practice styles (Chilingerian and Sherman 1997; Ozcan 1998). Getting physicians to see how their practice behavior ranks in relation to their peers is a step toward changing the culture of medicine and offering insights into a theory of clinical production management (Chilingerian and Glavin 1994).

In 1989, Chilingerian used DEA to investigate the nontechnical factors associated with clinical efficiency of 12 acute hospital surgeons and 24 acute hospital internists. Since clinical efficiency assumes a constant quality outcome,

the study classified each physician's patients into two outcomes: (1) satisfactory cases – i.e., patients discharged alive without morbidity; and (2) unsatisfactory cases – i.e., patients who experienced either morbidity, or who died in the hospital. Relying on a two output (low and high severity discharges with satisfactory outcomes), two input (ancillary service, and total length of stay) model, each of the 36 physicians were evaluated (Chilingerian 1989).

The study reported that 13 physicians were on the best practice frontier (i.e., 13 physicians produced good outcomes with fewer resources). After controlling for case-mix complexity and the severity mix of the patients, the DEA scores could not be explained by the type of patients treated. Moreover, younger physicians (age 40 and under) who belonged to a group practice HMO were more likely to practice efficiently.

In 1995, Chilingerian reported a 6-month follow-up study on the 36 acute care physicians (Chilingerian 1994). A Tobit analysis revealed that physicians affiliated with a group practice HMO were likely to be more efficient as well as physicians whose medical practices focused on a narrow range of diagnoses. Using DEA to investigate the most productive scale size of hospital-based physicians, the study reported that on average one high-severity patient utilized five times the resources of a low-severity patient.

In 1997, Chilingerian and Sherman evaluated the practice patterns of 326 primary care physicians in an HMO. They employed a seven output (gender and age-adjusted patients) eight input (acute hospital days, ambulatory surgery, office visits, subspecialty referrals, mental health visits, therapy units, tests, and emergency room visits) model. They demonstrated that the HMO's belief that generalists are more efficient than subspecialists was not supported. Using clinical directives to establish a cone ratio model, practice styles were identified that should have increased their office visits and reduced hospitals days. Thus, a DEA-based definition of efficiency identified relatively "efficient" physicians who could change their styles of practice by seeing more of their patients in their offices rather than in the hospital.

In 1998, Ozcan studied 160 primary care physicians' practice styles using DEA with weight restrictions to define preferred practice styles for a single medical condition. The study identified three outputs (low, medium, and high severity patients with otitis media, a disease that affects hearing) and five inputs (primary care visits, specialist visits, inpatient days, drugs, and lab tests). The study found 46 efficient and 114 less efficient practices. After defining preferred practice styles, even the efficient physicians could have a reduction in patient care costs of up to 24%.

16.2.5 Data Envelopment Analysis Versus Stochastic Frontier Analysis

Several health-care papers have compared DEA results with other techniques such as stochastic frontier analysis (SFA). These techniques use two vastly different

optimization principles. SFA has one overall optimization across all the observations to arrive at the estimates of inefficiency. DEA runs a separate optimization for each hospital, thus allowing a better fit to each observation and a better basis for identifying sources of inefficiency for each hospital.

The differences in optimizing principles used in DEA and regression estimates suggest that one might be preferred over the other to help solve different problems. For example, SFR may be more helpful to understand the future behavior of the entire population of hospitals. DEA might be used when the policy problem centers on individual hospitals how specific inefficiency can be eliminated. One problem with comparing DEA to SFA is that SFA requires a specification such as a translog function with a single dependent variable. Forcing DEA to utilize the same dependent variable as a single output will lead to the conclusion that the results are similar, but not exactly the same (see Bryce et al. 2000).

16.2.6 Reviewer Comments on the Usefulness of DEA

As the life cycle of DEA matures, we have seen an increase in the health applications published. However, the skeptics remain. The most skeptical view on frontier estimation has come from Joseph P. Newhouse at the Harvard School of Public Health. In 1994, in the *Journal of Health Economics*, Newhouse, raises several fundamental questions about frontier studies (Newhouse 1994). For example, he asks if one could define a frontier with certainty what purpose would it serve? Can frontier studies be used to set reimbursement rates? Can frontier analysis create benchmarks?

Unlike kilowatt-hours, Newhouse also suggests that because the product in health care is neither homogeneous, nor unidimensional the output problem is serious. He also identifies three other problems with modeling medical care delivery as a production process. First, frontier studies often omit critical inputs such as physicians, contract nurses, capital inputs, students, and researchers. Employing case-mix measures such as Diagnostic Related Groups, or Ambulatory Cost Groups, can also hide severity within a diagnosis or illness. He points out that DEA assumes no random error or random fluctuations, which is often not the case in health care.

The problem of measurement errors, or unclear, ambiguous, and omitted inputs and outputs will haunt every DEA health-care application. The threshold question is this – how seriously do these issues affect the results? While Newhouse claims a serious distortion, the question is largely empirical. Researchers must put that issue to rest each time they conduct a DEA study. Do the results make sense?

To advance the field, every DEA health-care study should test the following hypothesis:

Ho: Variations in DEA scores (i.e., excess utilization of clinical resources) can be explained by complexity, severity, and type of illness.

Though researchers cannot prove that DEA scores are measuring efficiency, they can disconfirm the case-mix hypothesis. If DEA scores are not associated with severity, patient characteristics or poor quality the DEA results become very interesting. One analytic strategy for disconfirming the case mix/quality hypothesis is to regress the DEA scores against the best available measures of case mix and patient characteristics. We discuss how to use Tobit models to validate DEA scores in another section of this chapter.

16.2.7 Summary

One conclusion to be reached from all of the health-care studies is that even when it appears that a substantial amount or resources could be "saved" if every hospital, nursing, or physician were as good as those who use medical techniques the least, DEA scores must always be interpreted with care. Researchers should assume that there is unaccounted case mix until that issue is put to rest.

In summary, there are two obstacles to advancing applications of DEA to health services research. The first impediment is the confusion around modeling hospitals, physicians, nursing homes, and other health-care providers. The findings from DEA studies lose credibility when inputs and outputs are defined differently from study to study. A second impediment that interferes with the development of generalizations from DEA applications is the lack of stability in the results from different studies. A variable such as quality of care, associated with productive efficiency in one study, disappears in another study (see Kooreman 1994; Rosko et al. 1995). Rarely have two researchers studied the same problem and when they have, rarely have they employed the same categories of inputs and outputs.

One way to deal with these difficulties is to try to ascertain how much of the efficiency scores is explained by case mix. For example, studies that blend DEA with Tobit or other statistical models can sharpen the analysis of best practices. If adequate controls are included in the model, the challenge of output heterogeneity in health care can be laid to rest (see Rosko et al. 1995; Kooreman 1994; Chilingerian 1995).

This review suggests that DEA is *capable* of producing new knowledge advancing the science of health-care management. As Chap. 1 has argued:

> . . .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 are required in linear and nonlinear regression models. . .. (See Chapter 1 of his handbook).

Before DEA becomes a primary tool to help policy makers and practicing clinical managers, researchers must shift from health care "illustrations of DEA" to advancing the field of performance improvement in the delivery of health services. If DEA is to reach its potential for offering new insights and making

a difference to clinicians and patients, DEA research has to overcome the many pitfalls in health care. The remainder of this chapter identifies the issues, and offers some practical suggestions for dealing with them. We conclude with a new approach that investigates quality frontiers in health care.

16.3 Health-Care Models

Medical care production is different from manufacturing. A traditional factory physically transforms raw materials into finished products. The customer is absent, so there is no participation or co-production. If demand is higher than capacity, inventory can be stored. In manufacturing, quality can be built into design of a product. Since goods can be inspected and fixed at every workstation, quality can improve productive efficiency.

In health care after patients are admitted to a care facility (or visit a clinic) there are three major clinical processes: (1) investigation/diagnosis, (2) treatment/therapy, and recovery. Health-care services are intangible "performances" that can only be experienced or used (Teboul 2002). They must be consumed immediately following production, and they cannot be stockpiled. Unused capacity – i.e., nursing care, empty beds, idle therapists, and the like – is a source of inefficiency. However, overutilization – i.e., unnecessary tests, X-rays, and surgeries, or unnecessary days in the intensive care unit or nursing home – is also a source of inefficiency. Once delivered, clinical services cannot be taken back and corrected patient perceptions are immediate. In the next section, efficiency in the health-care context is defined and discussed.

16.3.1 Clinical Efficiency Definitions

Clinical inefficiency in the provision of health-care services occurs when a provider uses a relatively excessive quantity of clinical inputs when compared with providers treating a similar case load and mix of patients. There are many categories of efficiency – technical, scale, allocative, and overall efficiency. Most studies in health care have measured the overall technical and scale components of clinical efficiency, which is represented by the "average productivity attainable at the most productive scale size. . .." (Banker, Charnes, and Cooper 1984, p. 1088). Most simply, technical inefficiency refers to the extent to which a DMU fails to produce maximum output from its chosen combination of factor inputs, and scale inefficiency refers to suboptimal activity levels.

Evaluating a health-care provider's clinical efficiency requires an ability to find "best practices" – i.e., the minimum set of inputs to produce a successfully treated patient. Technical inefficiency occurs when a provider uses a relatively excessive quantity of clinical resources (inputs) when compared with providers practicing

with a similar size and mix of patients. Scale inefficiency occurs when a provider is operating at a suboptimal activity level – i.e., the unit is not diagnosing and/or treating the most productive quantity of patients of a given case mix. Hence, hospital providers will be considered 100% efficient if they cared for patients with fewer days of stay and ancillary services and at an efficient scale size. Primary care providers will be considered efficient if they cared for their panels of patients with fewer visits, ancillary tests, therapies, hospital days, drugs, and subspecialty consults.

Researchers can use a variety of DEA models to measure and explain overall technical and scale efficiency. The CCR model, initially proposed by Charnes et al. (1978) is considered a sensitive model for finding inefficiencies (Golany and Roll 1989). In 1984, Banker et al. added another very useful model (BCC model) for health-care studies. The BCC model can be used to separate technical from scale efficiency. Both models (if formulated as input-minimizing) can be used to explore some of the underlying reasons for inefficiency – e.g., to estimate divergence from most productive scale size and returns to scale (Banker et al. 1984). Consequently, DEA can yield theoretical insights into the managerial problems or decision choices that underlie efficient relationships: e.g., magnitude of slack, scale effects of certain outputs on the productivity of inputs, marginal rates of substitution and marginal rates of transformation and so on (Cooper et al. 2000b).

Once DEA rates a group of providers efficient and inefficient, researchers, managers and/or policy makers can use this information to benchmark best practice by constructing a theoretical production possibility set. Analysts or researchers could use the DEA linear programming formulations to estimate potential input savings (based on a proportional reduction of inputs). Analysts or researchers can use the ratios of the weights u_r and v_i to provide estimates of marginal rates of substitution and marginal rates of transformation of outputs, measured on a segment of the efficient frontier (Zhu 2000). Again, analysts or researchers could use the BCC model to evaluate returns to scale – i.e., in the case of physicians, the effects of a small versus large proportion of high severity cases. Furthermore, the production possibility set is used to estimate alternative outputs of low severity and high severity patients that can be offered clinical services having utilized a mix of clinical services (for a primary care example see Chilingerian 1995).

16.3.2 How to Model Health-Care Providers: Hospitals, Nursing Homes, Physicians

The threshold question for DEA is what type of production model to choose. Depending on the type of health-care organization, there are many ways of conceptualizing the inputs and outputs of production. Since the selection of inputs and outputs often drives the DEA results, it is important to develop a

justification for selecting inputs and outputs. In the next section, we review DEA models used in various health applications. First, we differentiate clinical from managerial efficiency.

16.3.3 Managerial and Clinical Efficiency Models

In health care, technical efficiency is not always synonymous with managerial efficiency. On the one hand, technical efficiency in nursing homes, rehabilitation hospitals, and mental health facilities can be equated with managerial efficiency. On the other hand, medical care services especially in acute and primary care settings are fundamentally different in that there are two medical care production processes, and consequently types of technical and scale efficiency: managerial and clinical. Managerial efficiency requires practice management – i.e., achieving a maximum output from the resources allocated to each service department, given clinical technologies. Clinical efficiency requires patient management – i.e., physician decision-making that utilizes a minimal quantity of clinical resources to achieve a constant quality outcome, when caring for patients with similar diagnostic complexity and severity.

In the case of acute hospitals, the role of the manager is to set up and manage a decision-making organization whose basic function is to have clinical services ready for physicians and other clinical providers. For example, hospital managers must have admitting departments, dietary and diagnostic departments, operating rooms, and ICU and regular beds staffed and ready to go. Physicians make decisions to use these intermediate products and services, and managers make it all available. The major challenge facing hospital managers is to decide how much reserve capacity is reasonable given fluctuating patient admissions (daily and seasonal), and stochastic emergency events. Managerial efficiency can be equated with producing nursing care, diagnostic and therapeutic services, and treatment programs of satisfactory quality, using the least resources. Good practice management achieves managerial efficiency.

Physicians are fundamentally different from other caregivers. Not only are they providers of medical care but they also enjoy the primary "decision rights" for patient care, with little interference from management. They are the DMUs that steer a patient through various phases of patient care – such as office visits, primary care and diagnostic services, hospital admissions, consultations with other physicians, surgeries, drugs, discharges, and follow-up visits. Physicians organize and direct the entire production process, drawing on the talents of a hidden network of providers – nurses, therapists, dietary specialists, and the like. The reason that physician practice patterns are of interest is that 80–90% of the health-care expenditures in every system are the result of physician decision-making. These are dominant and highly influential DMUs. Clinical inefficiency then, as it is used here, refers to physicians who utilize a relatively excessive quantity of

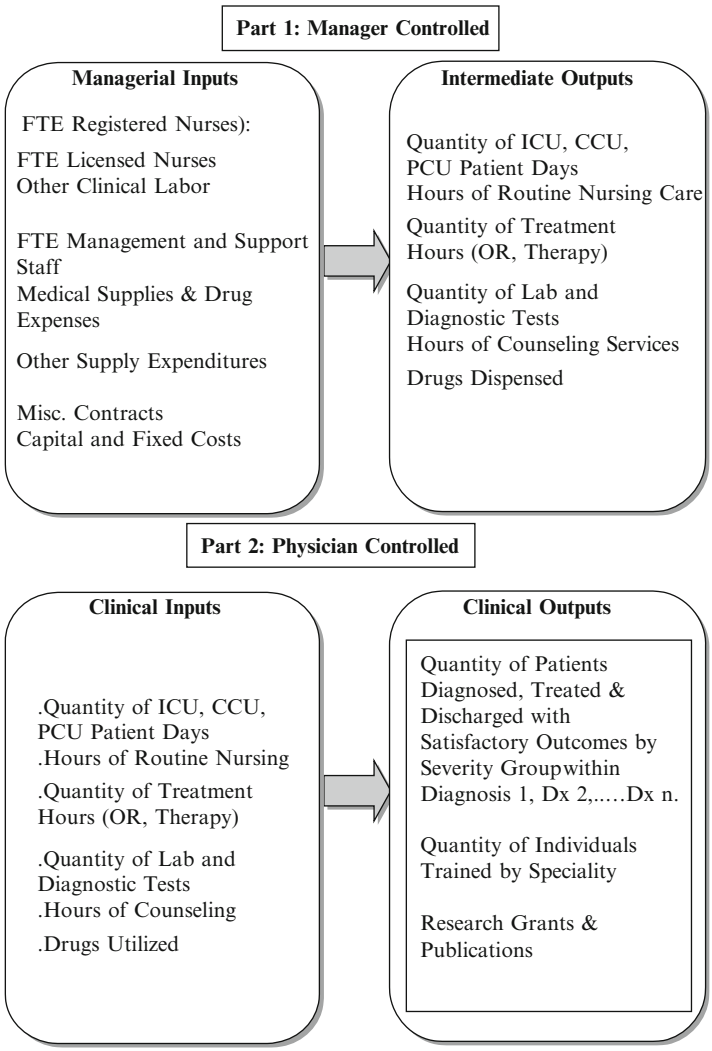


Fig. 16.1 Acute hospital as two-part DEA model (Chilingerian and Sherman 1990)

clinical resources, or inputs, when compared with physicians practicing with a similar size and mix of patients.

Figure 16.1 describes the medical service production system as a two-part service process: (1) a manager-controlled DMU, and (2) a physician-controlled DMU. In the diagram below, the intermediate outputs of the manager-controlled production process become the clinical inputs for the physician-controlled production process. A discharged patient is the final product, and the clinical inputs are the bundle of intermediate services that the patient received.

Hospital managers set up and manage the assets of the hospital. They control the labor, the medical supplies, and all expenditures related to nursing care, intensive care, emergency care, and ancillary services (such as lab tests, radiology, and other diagnostic services), pharmacy, dietary, as well as laundry, central supplies, billing, and other back-office functions. However, these departments (or functions) merely produce intermediate services that are available for utilization by physicians (see Chilingerian and Glavin 1994; Chilingerian and Sherman 1990; Fetter and Freeman 1986). Physician decision-making determines how efficiently these assets are utilized. Once a patient is admitted to hospital, physicians decide on the care program – i.e., the mix of diagnostic services and treatments, as well as the location and intensity of nursing care, and the trajectory of the patient. Physicians decide how and when to utilize nursing care, intensive care, emergency care, ancillary services, and other clinical inputs.

The productive efficiency of the hospital is complicated. A hospital can be clinically efficient, but not managerially efficient. A hospital can be managerially efficient, but not clinically efficient. More often, both parts of the production process are inefficient. If physicians over utilize hospital services the cost-per-patient day, cost-per-nursing hour, cost-per-test is reduced and the hospital appears to making the best use of its inputs.

To be efficient, clinical and nonclinical managers must perform two tasks very well. Clinical managers must manage physicians' decision-making (i.e., patient management) and nonclinical managers make the best use of all hospital assets by managing operations (i.e., practice management). Therefore, patient and practice management require an extraordinary amount of coordination and commitment to performance improvement.

16.3.3.1 Medical Center and Acute Hospital Models: Examples of Managerial Efficiency

Three examples of acute production models appear below. The first model (Burgess et al. 1998) includes five types of labor inputs and weighted beds as a proxy for capital, but excludes drugs, medical supplies and other operating expenses. The second model (Sexton et al. 1989) collapses nurses into one category, but adds physicians and residents and excludes beds. The third model collapses labor into one variable, includes other operating expenses and beds, but also adds a proxy measure of capital based on a count of the number of specialty and diagnostic services (Fig. 16.2–16.4).

The outputs are different in all three models. Conceptually if the inputs are costs, then the input/output ratios are cost per case, cost per procedure, cost per visit, or cost per nursing day. If the inputs are beds or FTE labor, then the input–output ratio is represented by labor utilized per admission, labor utilized per patient day, labor per surgery, and the like. Although mixing managerial inputs with clinical outputs is acceptable, managerial and clinical inefficiencies become indistinguishable.

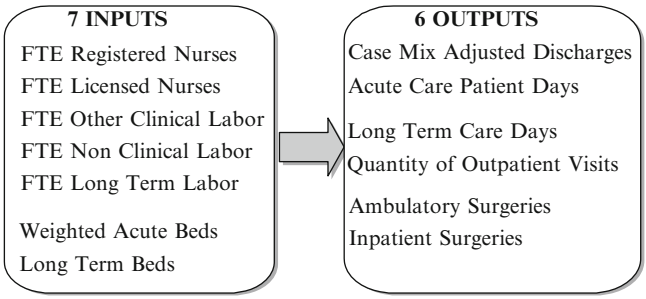


Fig. 16.2 Variables in general acute hospital model (Burgess et al. 1998)

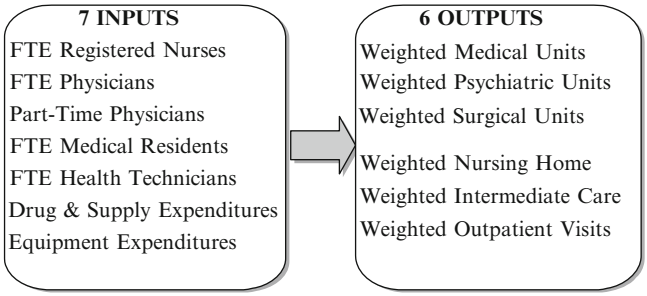


Fig. 16.3 Variables in a medical center study (Sexton et al. 1989)

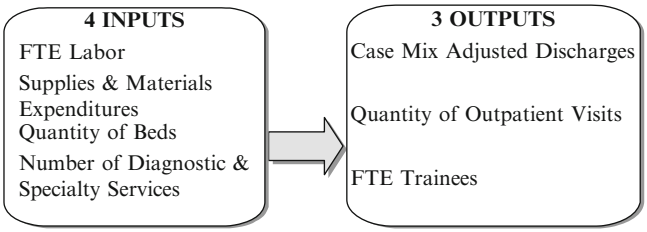


Fig. 16.4 Variables in urban hospital model (Ozcan and Luke 1992)

16.3.3.2 Nursing Homes: Another Example of Managerial Efficiency

Nursing home studies in the USA typically segment the outputs by sources of payments: quantity of residents supported by the state, or people without insurance. Figure 16.5 displays the inputs and outputs often used in DEA nursing home studies. The nine resource inputs are full-time equivalent (FTE) registered nurses, FTE licensed practical nurses, and FTE nurse aides, FTE other labor, and medical supplies and drugs, clinical and other supplies, and claimed fixed

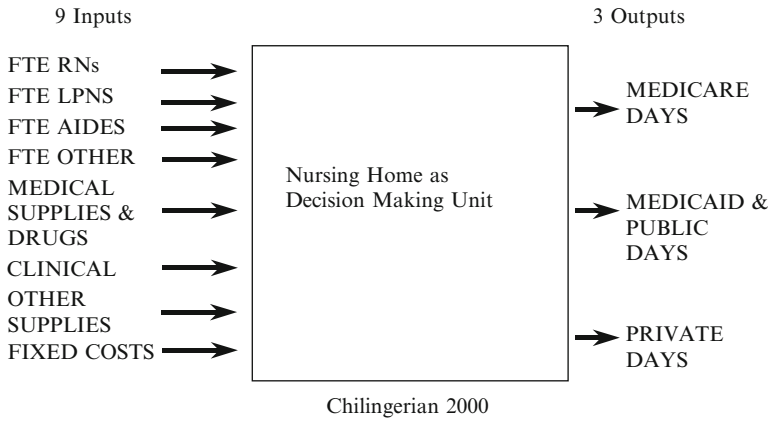


Fig. 16.5 Nursing home inputs and outputs (Chilingerian 2000)

costs (a proxy for capital). Since DEA can handle incommensurable data, the FTEs are in quantities, and the supplies, drugs, and fixed costs are measured by the amount of dollars spent. The outputs are the quantity of nursing home days produced during a given time period. In the Fig. 16.5, the outputs are the quantity of resident days broken into three payment classification groups: Medicare patients (a national program to pay for elderly care), Medicaid patients (a state program to pay for impoverished residents), and Private patients (residents without financial assistance).

Problems arise when the outputs are not homogeneous due to unaccounted case mix. For example, if the nursing home has skilled nursing beds, an Alzheimer unit, or a large proportion of patients older than 85, confused, requiring feeding, bathing, and toilet assistance, then the model is not measuring differences in pure productive efficiency. One solution is to collect information on patients' characteristics and regress the DEA scores against the patient, environmental, and managerial factors to be sure that the DEA scores are not due to case mix. For an example of this type of study, see Rosko et al. 1995.

16.3.3.3 Primary Care Physician Models: An Example of Clinical Efficiency

With growth of managed care, the primary care physician has emerged as an important force in the struggle for efficient and effective medical care. Since lab and radiology tests, prescription drugs, surgeries, and referrals to hospitals all require a physician's approval, physicians report cards or profiles have become a way to benchmark physician practice patterns. Managers could use DEA as a tool to profile and evaluate physicians.

Previous research has found that three patient variables drive managed care costs. They are patients' age, gender, and geographic location. Consequently, managed

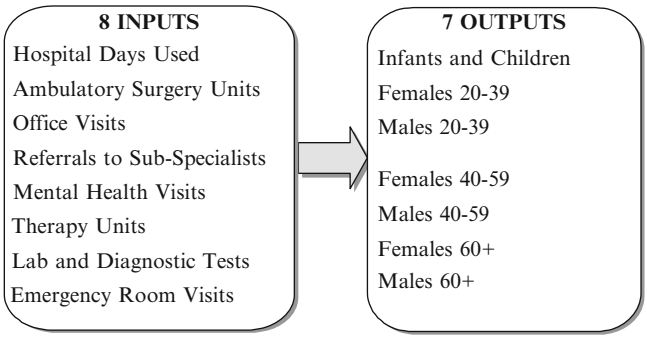


Fig. 16.6 Variables in a primary care physician study (Chilingerian and Sherman 1997)

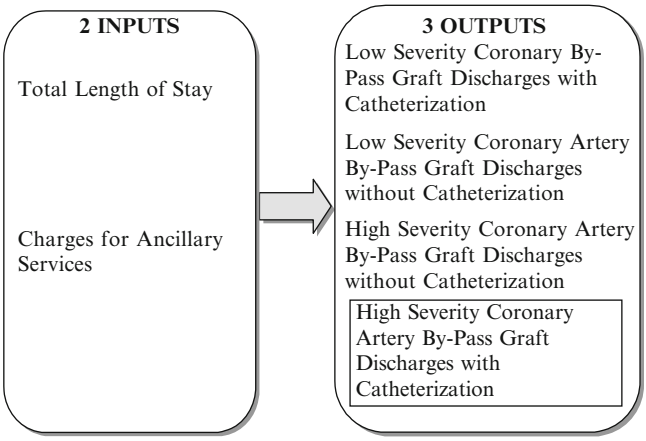


Fig. 16.7 Variables in a cardiac surgeon model of DRG 106 and 107 (Chilingerian et al. 2002)

care organizations set their budgets and prices based on these variables. The final product produced by physicians in managed care organizations is 1 year of comprehensive care for their patients. To care for patients, primary care physicians utilize office visits, hospital days, lab tests, and therapy units. Figure 16.6 is an example of how one large HMO conceptualized a physician DEA application.

16.3.4 Hospital Physician Models: Another Example of Clinical Efficiency

To study the practice behavior at the individual physician-level, analysts and researchers can use a variety of DEA models. For example, one recent study a study of 120 cardiac surgeons evaluated how efficiently they performed 30,000 coronary artery bypass grafts (CABG) on patients over a 2-year period (see Chilingerian et al. 2002). Figure 16.7 illustrates a two-input, four-output clinical production model used in that study. The two inputs are defined as

(1) the total length of stay (days) for the CABG cases handled and (2) the total ancillary and other charges (dollars) for the CABG cases handled. The ancillary and other charges input category includes ancillary, drug, equipment, and miscellaneous charges. The first input, length of stay, represents a measure of the duration of CABG admissions and the utilization of clinical inputs such as nursing care and support services. The second input, ancillary and other charges, represents a measure of the intensity of CABG admissions and the utilization of operating rooms, laboratory and radiological testing, drugs, and so on.

The four classes of clinical outputs represent completed CABG surgery cases. Since patients with more severe clinical conditions will likely require the use of more clinical inputs, the efficiency analysis must account for variations in case mix to be fair to surgeons or hospitals treating relatively sicker CABG patient populations. Accordingly, the outputs are defined by diagnostic category and severity level within diagnosis. In this example, a system of case mix classification called Diagnostic Related Groups (DRG) are used to segment outputs by complexity; moreover, a severity system called MEDSGRPS was used to further segment each DRG into low and high severity categories. The researchers treated DRG 106 and DRG 107 as separate clinical outputs because a CABG procedure with catheterization is more complicated and requires more clinical resources. As explained above, each DRG was further divided into low-severity and high-severity cases.

16.3.5 Profitability Models: A Nursing Home Example

Although "profit" is still a dirty word in health care, there is a need to do more performance studies that look at revenue and expenses, and investigate the factors affecting profitability. In these studies, the maximum profit includes actual profit, plus maximum overall inefficiency (see Cooper et al. 2000a). Figure 16.8 illustrates an example of a profitability model for a nursing home.

Since the performance measure, takes the form of Profit = Revenues−Expenses, an additive mode could be used (see early chapters in this handbook). The additive model shown below has several advantages.

$$\max z = \sum_{r=1}^{s} \mu_r y_{ro} - \sum_{i=1}^{m} v_i x_{io} + u_o,$$

subject to

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

$$\mu_r, v_i \geq 1.$$

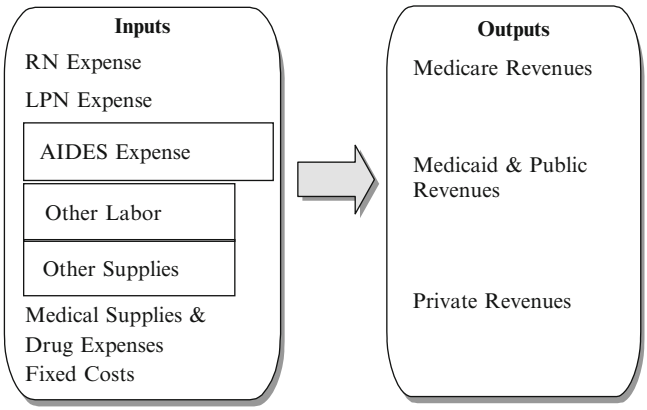


Fig. 16.8 Hypothetical example of nursing home inputs and outputs in the profitability model

One advantage of the additive model is that the objective function, max z, can be interpreted as maximizing profit, or maximizing an excess of revenue over expenses. The model contains an unconstrained variable, $u_{o,}$ which forces the model to assume variable return to scale (see early chapters in this handbook). Given the advantages and disadvantages of having a great deal of capacity, this assumption works for a profitability model. The optimal value of v_r^* can provide insights with respect to the trade-offs associated with adding skilled nursing, or reducing other labor or fixed expenses.

Employing profitability models to evaluate health-care organizations is a promising area for research for two reasons. First, profit models make the evaluation of the complex delivery systems such as acute care hospitals more manageable. Second, this approach opens up the opportunity to use a multi-stage evaluation approach. With a three-stage approach, DEA could be used to investigate the underlying reasons for the profitability. During stage one, an average DEA score of the clinical efficiency of the each hospital's busiest physicians could be obtained by studying the individual physician level. For the second stage, a profitability model would be run, obtaining DEA scores for each hospital in the comparison set. Finally, the third stage would regress the hospital's profitability scores against the explanatory variables, including the average clinical efficiency of physicians.

The variables that influence health-care profit margins include the following: cost drivers, prices (relative to competitors), relative scale and capacity, the market share, quality of care, and technical delivery mode (Neuman et al. 1988). According to the literature, the main cost drivers are: the volume of patients, within diagnostic categories, adjusted by severity; the practice style and clinical behavior of the medical staff; and the degree of managerial and clinical efficiency (Chilingerian and Glavin 1994).

16.4 Special Issues for Health Applications

There are conceptual, methodological, and practical problems associated with evaluating health-care performance with DEA. First, conceptualizing clinical performance involves identifying appropriate inputs and outputs. Selecting inputs and outputs raises several questions – Which inputs and outputs should the unit be held accountable? What is the product of a health-care provider? Can outputs be defined while holding quality constant? Should intermediate and final products be evaluated separately?

Another conceptual challenge involves specifying the technical relationship among inputs. In analyzing clinical processes, it is possible to substitute inputs, ancillary services for routine care days, more primary care prevention for acute hospital services, and the like. Within the boundaries of current professional knowledge, there are varieties of best practices. Consequently, an evaluation model should distinguish best practices from alternative practice styles.

Methodological problems also exist around valid comparisons. Teaching hospitals, medical centers, and community hospitals are the same. Nursing homes and skilled nursing facilities may not be the same. These institutions have a different mix of patients and have different missions. Finally, the information needed to perform a good study may not be available. Much of the interesting DEA research has not relied on an easily accessible dataset, but required merging several data bases (see Burgess et al. 1998; Rosko et al. 1995).

There are problems about choice of inputs and outputs, and especially finding an "acceptable" concept of product/service. Which inputs and outputs should physicians be held accountable? In addition, there are other issues about measures and concepts. For example:

- Defining Models from Stakeholder Views
- Selection of Inputs and outputs
- Should inputs include environmental and organizational factors?
- Problems on the Best Practice Frontier
  - Are the input factors in medical services substitutable?
  - Are returns to scale constant or variable?
  - Do economies of scale and scope exist?

We discuss each of these issues in the next section.

16.4.1 Defining Models from Stakeholder Views

The models used in DEA depend on the type of payment system and the stakeholder perspective. Every health-care system faces different payment and financing schemes that determine the location of financial risk and, subsequently, the interests of the stakeholders. Examples of payment and finance systems include

fee-for-service reimbursement, prospective payment, fixed budgets, and capitation. All of these payment systems have one thing in common – they are all seriously flawed. They can imperil quality and often have contributed to increased health-care costs.

There are also many stakeholders, each with different and sometimes conflicting interests. For example, there are institutional providers (hospitals, nursing homes, clinics, etc.) as well as individual providers (physicians, nurses, therapists, etc.), governments (local, regional, and national health authorities), third party payers (insurance companies, sick funds, etc.), and patients and their families. All stakeholders want better quality and better access for patients. Although stakeholders talk about more efficient care, they focus most of their attention on how to stem their rising costs.

In many countries, ambulatory physician services are reimbursed on a fee-for-service basis. Since there is no incentive to reduce patient care, from a provider, self-interest standpoint under an unrestricted fee-for-service the care philosophy becomes *more care = better care*, which can lead to over utilization of clinical services and rising health costs. Unless physician decision-making is managed by strong professional norms and benchmarking best practices, systems that create incentives to maximize patient visits, hospital days and procedures leads to rising costs and less efficient care.

Prepayment is the flip side of fee-for-service. With a prepayment system, there is a contractually determined, fixed price/payment for a defined set of services. For example, countries such as Germany, Italy, and the USA have prospective payment systems that reimburse the hospital for each patient admission based on diagnosis irrespective of the actual costs. Taking the hospital's perspective, to enhance revenue the hospital must keep the beds full (increase admissions), and minimize the hospital's cost-per-admission, which of course can lead to over utilization of clinical services and rising health costs.

Capitation and gate keeping is another payment scheme that places physicians' interests at risk. The idea behind this design is that by having the primary care physician (PCP) assume responsibility for all aspects of care for a panel of patients, the patients will enjoy a greater continuity of care while the HMO achieves a more efficient care process at a more consistent level of quality. The "gatekeeper" receives a fixed monthly payment for each patient (adjusted actuarially by age and gender) as an advance payment to provide all the primary care the patient needs. The incentives can be designed so that primary care physicians prosper if they keep their patients out of the hospital and away from specialists. However, the less care the PCPs provide, the greater is the financial gain. Failure to provide needed care in a timely manner to HMO patients may make these patients' treatments more expensive when their illnesses become more advanced. This type of payment scheme can lead to underserved patients and rising health costs.

The payment systems described above can create weak or even distorted motivating environments. DEA models should be defined from a particular stakeholder point-of-view; moreover, models should be selected to ensure that patients are neither underserved nor overserved.

16.4.2 Selecting Appropriate Health-Care Outputs and Inputs: The Greatest Challenge for DEA

Acute care hospitals and medical centers are complex medical care production processes, often bundling hundreds of intermediate products and services to care for each patient (Harris 1979). Major surgery might also require major anesthesia, blood bank services, hematology tests, pathology and cytology specimen, drugs, physical therapy, an intensive care bed, and time in routine care beds. Given the complexity, the greatest challenge for DEA health-care applications lies in conceptualizing and measuring the inputs and outputs (see Chilingerian and Sherman 1990; Newhouse 1994).

Dazzling new technologies, desires to improve efficiency, and new consumer attitudes have changed service capacities, practice behaviors and outcomes. The explosive growth in outpatient surgeries is a good example. In the USA, in 1984 only 400,000 outpatient surgeries were performed; by 2000 that number grew to 8.3 million (Lapetina and Armstrong 2002). The intensity and mix of outpatient and inpatient surgeries are continually changing. When outpatient surgeries shift from tooth extractions to heterogeneous procedures such as hernia repairs, cataract and knee surgeries, and noninvasive interventions, simple counts of the quantity of surgical outputs are misleading. The mix and type of surgeries must be taken into account as well as the resources and technological capabilities of the DMUs.

The first DEA paper published by Nunamaker (1983) used a one input–three-output model. The outputs used crude case-mix adjustments (pediatric days, routine days, and maternity days). The single input was an aggregate measure of inpatient routine costs, which will lead to unstable results unless the inputs among the comparison set homogeneous. We know that hospital salaries, education, experience and mix of the nurses and other staff, the vintage of the capital and equipment, the number of intensive beds, medical supplies and materials are not the same (see Lewin 1983).

To make the DEA results useful for practice, it makes sense to segment inputs into a few familiar managerial categories that make up a large percent of the expenditures. Since acute hospitals have high labor costs, it makes sense to distinguish the types of personnel: i.e., the number of full time equivalent (FTE) registered nurses, FTE licensed practical nurses, FTE nurses aids, FTE therapists, other FTE clinical, and general administrative staff. Selecting managerially relevant categories is necessary if analysts are to use DEA to identify the sources waste and inefficiency.

16.4.2.1 Take Two Aspirin and Call Me in the Morning

With respect to outputs, there are two requirements. First, health-care outputs cannot be adequately defined without measures of case-mix complexity and severity. Second, it makes no sense to evaluate the efficiency of a medical service that

results in an adverse event: such as morbidity, mortality, readmissions, and the like. To construct the "best practices" production frontier, observed behavior is evaluated as clinically inefficient if it is possible to decrease any input without increasing any other input and without decreasing output; and if it is possible to increase any output without decreasing any other output and without increasing any input (see Chap. 1 in this handbook). A DMU will be characterized as perfectly efficient only when both criteria are satisfied.

If hospital discharges include both satisfactory and unsatisfactory outcomes, a hospital could be considered efficient if they produced more output per unit of input. If a patient die in the operating room and few clinical resources are utilized, the outcome falls short of the clinical objective. Clinical efficiency requires that outcomes be considered in the performance standard. Therefore, as a concept, clinical efficiency makes sense if and only if the clinical outputs achieve constant quality outcomes. Therefore, the best attainable position for a health-care organization is when a unit achieves maximum the outputs should guarantee constant quality outcomes. Unsatisfactory outcomes should be taken out of the DEA analysis and evaluated separately (see Chilingerian and Sherman 1990). Likewise, if a primary care physician uses fewer office visits, fewer hospitals days, and fewer tests because she or he are postponing care, they should not be on the best practice frontier.

16.4.2.2 Using DEA to Adjust Outputs for Patient Characteristics and Case mix

If we are to go beyond simple illustrations, a DEA study should provide some guarantee that the outputs are similar. For example, it has long been established that acute patient care cannot be measured by routine patient days, maternity days, and pediatric days alone because patient days are an intermediate output and a poor indicator for other services such as blood work, X-rays, drugs, intensive care days, physical therapy and the like. Therefore, it is important that both clinical and ancillary services be included. In contrast to acute hospitals, nursing home studies that define nursing days segmented by payer-mix and/or age-adjusted may be an acceptable proxy of final outputs.

To help guide policy makers or practitioners, researchers might consider the following four-part DEA analytic strategy. The first part would begin by running the some DEA models and the second part by regressing the DEA scores against the case-mix and patient characteristic variables using a censored regression model such as Tobit. If the goodness-of-fit test is significant, adjust each health-care provider outputs by multiplying them by the ratio of the original DEA score to the Tobit's predicted DEA score. Some providers operating with a less complex case mix will have their outputs lowered and other providers operating with a more complex mix will have their outputs increased. A second DEA model would be run during the third part and the last step would be regressing the "new" efficiency scores again to validate that they are unrelated to any control variables other than the critical managerial or policy variables of interest. Using DEA with this analytic

strategy might provide some additional insights to policy makers or managers searching for answers to questions such as what is the impact of ownership structure or leadership on the productivity of the industry.

16.4.3 Should Environmental and Organizational Factors Be Used as Inputs?

Random variations in the economic environment (such as labor markets, accidents, epidemics, equipment failures, and weather) and organizational factors (such as leader behavior, employee know-how, and coordination techniques) can influence organizational performance. However, there are so critical inputs and outputs that must be included in health applications, that nondiscretionary variables should be omitted from the DEA model. For a strong health application, the clinical production model should only include resources that clinical decision-makers utilize and manage. Environmental (and other organizational) factors omitted from the DEA model should be included in a second stage model that investigates variables associated with the DEA scores.

Every researcher must decide on a reasonable conceptual model. To guide research, collaborating with practicing managers or policy makers to identify relevant environmental and explanatory factors can only strengthen a health application.

16.4.4 Problems on the Best Practice Frontier: A Physician Example[1]

As described in Chap. 1, of this book, DEA relies on the Pareto-Koopmans definition of efficiency. For example, to construct the "best practice" production frontier for primary care physicians, observed behavior is evaluated by using the following input–output criteria:

1. A physician is clinically inefficient if it is possible to utilize fewer clinical resources without increasing any other resources and without decreasing the number of patients cared for;
2. A physician is clinically inefficient if it is possible to care for more of a given segment of patients without decreasing the care given to any other segment of patients and without utilizing more clinical resources.

[1]The next section draws heavily on Chilingerian and Sherman (1995), Rosko (1990) and Charnes, Cooper and Rhodes (1978).

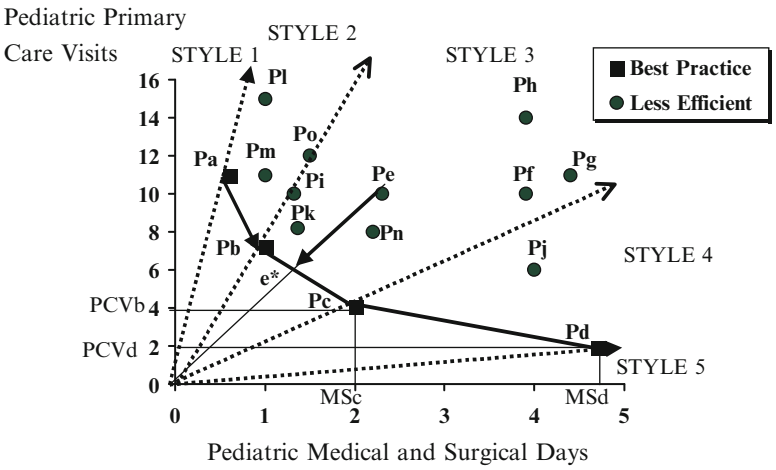


Fig. 16.9 Two-dimensional picture of pediatrician practice behavior

A physician can be characterized as clinically efficient only when both criteria are satisfied and when constant quality outcomes can be assumed.

A brief example with a small group of pediatricians illustrates how DEA can be used to define a best practice production frontier for primary care physicians. Consider 14 primary care pediatricians who cared for 1,000 female children 1–4 years old from a homogeneous socio-economic background. Figure 16.9 plots the amount of medical-surgical hospital days used and primary care visits for one full year, for the 1,000 children.

The Pareto-Koopmans efficiency criteria suggests that any pediatrician in Fig. 16.9 who was lower and to the left of another pediatrician was more successful because he or she used fewer clinical inputs to care for the "same" type of children. By floating a piecewise linear surface on top of the actual observations, Fig. 16.9 also plots a best practices production frontier (BPPF) for each of the 14 pediatricians. As shown in Fig. 16.9, pediatricians Pa, Pb, Pc, and Pd dominate Pe, Pf, Pg, Ph, Pi, Pj, Pk, Pl, and Pm. Accordingly, Pa, Pb, Pc, and Pd, physicians not dominated by any other physicians lie on a best practices production frontier and all designated as frontier points.

DEA defines a best practice frontier by constructing a set of piecewise linear curves, such that any point on the BPPF is a weighted sum of observed DMUs. DEA measures the relative efficiency of each physician by the relative distance of that physician from the frontier. Physicians on the frontier (points Pa, Pb, Pc, and Pd) have a DEA score of 1 and physicians off the frontier have a value between 0.0 and 1.0. For example, in Fig. 16.9, the efficiency score of Pe, a pediatrician less efficient than Pb and Pc, can be measured by constructing a line segment from the origin to point Pe. This line segment crosses the BPPF at point e'. The efficiency rating of Pe is equal to the ratio of the length of line segment Oe* to the length of OPe. Since the numerator Oe* is less than the denominator OPe, the efficiency rating of physician Pe will be less than 1.0.

The practice of pediatric primary care medicine is a very complicated production problem where there are many slightly different ways to mix inputs to care for a population of patients. To solve the problem, each physician adopts a "style of practice"[2] that can be represented by the mix of clinical inputs used. The geometry of convex cones provides a pictorial language for interpreting the practice style of physicians. In mathematics, a cone is a collection of points such that if a physician P is at any point in the collection, then μP is also in the collection for all real scalars $\mu \geq 0$ (Charnes et al. 1953). However, cones can also be interpreted as a linear partition of physician practice styles based on a set of linear constraints such as a range of substitution ratios.

For example, in Fig. 16.9 the ray that begins at the origin (0) and intersects the frontier at Pb and the ray that begins at the origin and intersects the frontier at Pc define the upper and lower boundaries of a cone or *practice style 3*. All points contained on the line that begins at the origin (0) and intersects the frontier at Pb have the same ratio of medicine and surgery days to primary care office visits. Therefore, a physician practicing *style 3* will have a ratio of visits to hospital days that falls between the line that begins at the origin (0) and intersects the frontier at Pb and the line that begins at the origin and intersects the frontier at Pc.[3]

So, each cone can also be thought of as a different practice style. For example, practice style 2 represents physicians who use relatively more visits than hospital days. In contrast, practice style 5 represents physicians who use relatively more medical-surgical days than primary care office visits. By modifying the standard DEA model, it can be used to locate cones or natural clustering of DMUs based on the optimal weights assigned by the linear program. The next section illustrates how the new DEA concept of assurance regions can be used to identify and investigate practice styles.

16.4.4.1 The Concept of a Preferred Practice Cone or Quality Assurance Region

The largest single expense item for an HMO is hospital days. Lower utilization rates for hospital days represent more desirable practice style. Physicians who succeed in substituting primary care visits and ambulatory surgeries for medical-surgical days will contribute to a more efficient delivery system. Understandably, most HMO managers would prefer their physicians to reduce hospital days and increase their primary care visits.

[2]Practice style, as it used here, refers to the treatment patterns defined by the specific mix of resource inputs used by a physician (or group of physicians) to care for a given mix of patients.

[3]Although the two-input single output problem can be solved graphically, multi-input multi-output problems require a mathematical formulation that can only be solved by using a linear programming model.

Microeconomics offers a way to quantify the clinical guideline that primary care physicians (PCPs) should trade-off medical surgical days (input X_2) for more primary care visits (input X_1) while maintaining a constant output rate. Economists refer to the change in medical-surgical days (ΔX_2) per unit change in primary care visits (ΔX_1) as the marginal rate of technical substitution – i.e., $-\mathrm{d}x_2/\mathrm{d}x_1$, which equals the marginal product of medical surgical days (ω_h) divided by the marginal product of primary care visits (ω_v).[4]

In Fig. 16.9 every physician on the BPPF practices an input minimizing style of care – i.e., their DEA efficiency score = 1. However, not all of these physicians on the BPPF are practicing minimum-cost care. In other words, the mix of inputs used by some of these physicians is not the most cost effective. Consider the physicians Pd in practice style 5 and Pc in practice style 4. Since style 4 cares for the same patients with fewer hospital days, cost conscious HMO managers would probably want their primary care physicians practicing to the left of style 5.

Now, consider the BPPF in Fig. 16.9. If the quantities of medical-surgical days are increased from MSc to MSd, the same panel of patients will be cared for with slightly fewer of the less expensive primary care visits and much more of the expensive medical and surgical days. This points out a weakness in using DEA to estimate the best practice production frontier. From an HMO's perspective, although the practice style of physician Pd represents an undesirable practice region, the standard DEA model would allow Pd to be on the BPPF.

Modified versions of additive DEA models allow managers to specify bounds on the ratio of inputs such as the ratio of hospital days to primary care office visits (see Cooper et al. 2000a). These bounds can be called a preferred practice cone (Chilingerian and Sherman 1997). This new development in DEA bounds the optimal weights (or marginal productivities) and narrows the set of technically efficient behaviors. By incorporating available managerial knowledge into a standard model, an assurance region truncates the BPPF frontier – i.e., tightens the production possibility set – and opens up the potential to find more inefficiency.

16.4.4.2 Constant Versus Variable Returns to Scale

One difficult issue with every health application is whether to use DEA models that assume variable or constant returns to scale. Researchers should address this question based on prior knowledge and logical inferences about the production context. While imaginative guesses are tolerable, it is unacceptable to pick a model to get "better looking" DEA results.

Although a health-care organization's production function may exhibit variable returns to scale, there are intuitive reasons for expecting that a physician's clinical

[4] In economics, holding constant all other inputs, the marginal product of an input is the addition to total output resulting from using the last unit of input. The ratio of the marginal products is the marginal rate of technical substitution, defined as $-dx_2/dx_1$.

production function exhibit constant returns to scale (Pauly 1980). Physicians are taught that similar patients with common conditions should be taken through the same clinical process. Thus if a surgeon operates on twice as many patients with simple inguinal hernias or performs twice as many coronary bypass grafts, they would expect to use twice as many clinical resources. Consequently, scaling up the quantity of patients should result in a doubling of the inputs.

Hospitals and physician practices both vary in size. In North America, the largest stand-alone hospitals have less than 1,500 beds. In Europe, there are medical centers with as many as 2,000 and 3,000 beds. What is the minimal efficient bed size or the most productive bed size? Some primary care physician practices vary from 1,200 to 4,000 patients. From a quality standpoint, what is the most optimal patient size? Questions about returns to scale should be addressed in future DEA research and would benefit from cross-cultural comparisons.

16.4.4.3 Scale and Scope Issues

Health-care providers and organizations are multiservice firms, offering many clinical services to provide convenient, one-stop shopping, to connect diagnostic services with treatments, etc. Is it more efficient to offer many services under one roof? If offering many services together is more efficient, then economies of scope exist.

On the contrary, the rise in "focused factories" in health care such as hernia clinics, heart clinics, hip replacement centers, and the like, is stirring a great deal of interest throughout the world. Are patients better off going to a super-focused clinic? For example, is a cataract hospital more efficient than a general hospital performing fewer cataract procedures in its operating rooms? Are the efficiency gains of a focus strategy large or small? Questions about scale versus scope should also be addressed in future DEA research.

16.4.5 Analyzing DEA Scores with Censored Regression Models

DEA's greatest potential contribution to health care is in helping managers, researchers, and policy makers understand why some providers perform better or worse than others do. The question can be framed as follows – How much of the variations in performance are due to: (1) the characteristics of the patients, (2) the practice styles of physicians, (3) the microprocesses of care, (4) the managerial practices of the delivery systems, or (5) other factors in the environment? The following general model has been used in this type of health-care study:

DEA Score = f (ownership, competitive pressure, regulatory pressure, demand patterns, wage rates, patient characteristics, physician or provider practice

characteristics, organizational setting, managerial practices, patient illness characteristics, and other control variables).

The DEA score depends on the selection of inputs and outputs. Hence every health application is obliged to disconfirm the hypothesis that DEA is not measuring efficiency, but is actually picking up differences in case mix or other nondiscretionary variables. The best way to validate or confirm variations in DEA scores is to regress the DEA scores against explanatory and control variables. But what type of regression models should be used?

If DEA scores are used in a two-stage regression analysis to explain efficiency, a model other than ordinary least squares (OLS) is required. Standard multiple regression assumes a normal and homoscedastic distribution of the disturbance and the dependent variable; however, in the case of a limited dependent variable the expected errors will not equal zero. Hence, standard regression will lead to a biased estimate (Maddala 1983). Logit models can be used if the DEA scores are converted to a binary variable – such as efficient/inefficient. However, converting the scores < 1 to a categorical variable results in the loss of valuable information; consequently logit is not recommended as a technique for exploring health-care problems with DEA.

Tobit models can also be used whenever there is a mass of observations at a limiting value. This works very well with DEA scores which contain both a limiting value (health-care providers whose DEA scores are clustered at 1) and some continuous parts (health-care providers whose DEA scores fall into a wide variation of strictly positive values < 1). No information is lost, and Tobit fits nicely with distribution of DEA scores as long as there are enough best practice providers. If, for example, in a sample of 200 providers less than 5 were on the frontier, a Tobit model would not be suitable.

In the econometrics literature, it is customary to refer to a distribution such as DEA as either a truncated or a censored normal distribution. There is, however, a basic distinction to be made between truncated and censored regressions. According to one source:

> The main difference arises from the fact that in the censored regression model the exogenous variables x_i are observed even for the observations for which $y_i > L_i$. In the truncated regression model, such observations are eliminated from the sample. (Maddala 1983:166)

Truncation occurs when there are no observations for either the dependent variable, y, or the explanatory variables, x. In contrast, a censored regression model has data on the explanatory variables, x, for all observations; however the values of the dependent variable are above (or below) a threshold are measured by a concentration of observations at a single value (Maddala 1983). The concentration of threshold values is often based on an actual measure of the dependent variable – i.e., zero arrests, zero expenditures – rather than an arbitrary value based on a lack of information.

DEA analysis does not exclude observations greater than 1, rather the analysis simply does not allow a DMU to be assigned a value greater than 1. Hence, Chilingerian (1995) has argued that DEA scores are best conceptualized as a

censored, rather than a truncated distribution. The censored model would take the following form:

Efficiency score = actual score if score < 1
Efficiency score = 1 otherwise

There is a substantial literature on modeling data with dependent variables whose distributions are similar to DEA scores. For example, empirical studies with the number of arrests after release from prison as the dependent variable (Witte 1980) or the number of extra marital affairs as the dependent variables (Fair 1978) are among the best known examples in the published literature on the Tobit censored model. Each of these studies analyzes a dependent variable censored at a single value (zero arrests, zero marital affairs) for a significant fraction of observations. Just as the women in Fair's study can do no better than have zero extra-marital affairs, neither could a relatively efficient health-care provider be more efficient than 1. Thus, one could equate zero extramarital affairs or zero arrests to a "best practicing" hospital or provider.

A censored Tobit model fits a line that allows for the possibility of hypothetical scores > 1. The output can be interpreted as "adjusted" efficiency scores based on a set of explanatory variables strongly associated with efficiency. To understand why censored regression models make sense here, one must consider how DEA evaluates relative efficiency.

DEA scores reflect relative efficiency within similar peer groups (i.e., within a "cone of similar DMUs") without reference to relative efficiency among peer groups (i.e., cones). For example, an efficient provider scoring 1 in a peer group using a different mix of inputs (i.e., rates of substitution) may produce more costly care than a provider scoring 1 in a peer group using another mix of inputs (Chilingerian and Sherman 1990). Superior efficiency may not be reflected in the DEA scores because the constraints in the model do not allow a DMU to be assigned a value greater than 1. If DEA scores could be readjusted to compare efficiencies among peer groups, some physicians could have a score that is likely to be greater than 1. Despite the advantages to blending nonparametric DEA with censored regression models in practice, some conceptual problems do arise.

The main difficulty of using Tobit to regress efficiency scores is that DEA does not exactly fit the theory of a censored distribution. The theory of a censored distribution argues that due to an underlying stochastic choice mechanism or due to a defect in the sample data there are values above (or below) a threshold that are not observed for some observations (Maddala 1983). As mentioned above, DEA does not produce a concentration of ones due to a defect in the sample data, rather it is embedded in the mathematical formulation of the model.

A second difficulty of using Tobit is that it opens up the possibility of rank ordering superior efficiency among physicians on the frontier – in other words "hypothetical" scores > 1. In production economics, the idea that some DMUs with DEA scores of 1 may possibly have scores > 1 makes no sense. It suggests that some candidates for technical efficiency (perhaps due to random shifts such as luck, or measurement error) are actually less efficient.

Despite these drawbacks, blending DEA with Tobit model's estimates can be informative. Although DEA does not fit the theory of a censored regression, it easily fits the Tobit model and makes use of the properties of a censored regression in practice. For example, the output can be used to adjust efficiency scores based on factors strongly associated with efficiency.

Tobit may have the potential to sharpen a DEA analysis when expert information on input prices or exemplary DMUs is not available. Thus, in a complex area such as physician utilization behavior, Tobit could help researchers to understand the need to introduce boundary conditions for the DEA model's virtual multipliers.

The distribution of DEA scores is never normally distributed, and often skewed. Taking the reciprocal of the efficiency scalar, (1/DEA score), helps to normalize the DEA distribution (Chilingerian 1995).

Greene (1983) points out that for computational reasons, a convenient normalization in Tobit studies is to assume a censoring point at zero. To put a health-care application into this form, the DEA scores can be transformed with the formula:

$$\text{Inefficiency score} = (1/\text{DEA score}) - 1$$

Thus, the DEA score can become a dependent variable that takes the following form:

$$\text{DEA Inefficiency score} = xB + u \text{ if efficiency score} > 0$$
$$\text{DEA Inefficiency Score} = 0 \text{ otherwise}$$

Once health-care providers' DEA scores have been transformed, Tobit becomes a very convenient and easy method to use for estimating efficiency. The slope coefficients of Tobit are interpreted as if they were an ordinary least squares regression. They represent the change in the dependent variable with respect to a one unit change in the independent variable, holding all else constant.

When using Tobit models they can tested with a log-likelihood ratio test. This statistic is calculated by $-2\log(\lambda)$, where $\log \lambda$ is the difference between the log of the maximized value of the likelihood function with all independent variables equal to zero, and the log of the maximized values of the likelihood function with the independent variables as observed in the regression. The log-likelihood ratio test has a chi square distribution, where the degrees of freedom are the number of explanatory variables in the regression.

16.5 New Directions: From Productive Efficiency Frontiers to Quality-Outcome Frontiers

Although most applications of DEA have been applied to estimations of technical efficiency and production frontiers, the methodology offers an empirical way to estimate quality frontiers. For example, any dimension of quality can be assessed by employing multiple indicators in a DEA model and comparing a provider against a composite unit projected onto a frontier.

Table 16.1 Nursing home outcomes: 100 patients traced for 6 months

Nursing home	Functional status Q_1	Functional status Q_3
N1	0.15	0.40
N2	0.15	0.14
N3	0.35	0.35
N4	0.30	0.15
N5	0.40	0.50
N6	0.20	0.30

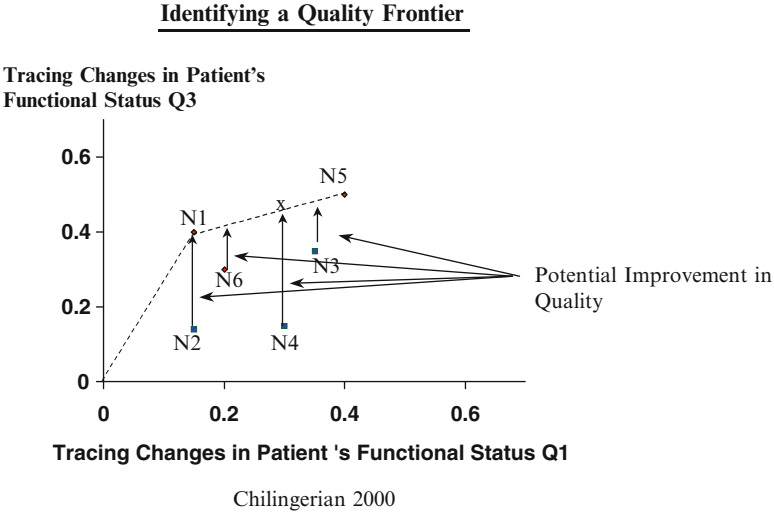


Fig. 16.10 Two-dimensional picture of a quality frontier

Although DEA could be applied to any type of quality evaluation, the following example is from a nursing home, demonstrating how to measure improvements in functional status. Drawing heavily on a paper by Chilingerian (2000), the following illustration explains how DEA works in estimating outcome frontiers. Consider a group of six nursing homes each with 100 patients whose bed mobility, eating, and toilet use have been traced for a period of 6 months.

Table 16.1 displays the overall functional status of the 100 patients in each of the six homes from quarter 1 to quarter 3. The functional status represents a vector of improvements or deteriorations of functional status in terms of the proportion of residents in quarter 1 completely independent and the changes to that group in quarter 3. That situation can be depicted graphically in Fig. 16.10 as a piecewise linear envelopment surface.

In Fig. 16.10, DEA identifies nursing homes that had the greatest improvement in functional quality over the 6-month period by allowing a line from the point of origin to connect the extreme observations in a "northwesterly" direction. Nursing homes N1 and N5 had the greatest improvement in the functional

independence of their patient populations and the dotted lines connecting N1 and N5 represent the best practice frontier.

The vertical lines above N2, N6, N4, and N3 represent the potential improvements in quality outcomes if N2, N4, N6, and N3 had performed as well as a point on the line between N1 and N5. Note that since the initial functional status is a nondiscretionary variable, the potential improvement can only occur along the vertical line. Associated with each underperforming nursing home is an optimal comparison point on the frontier that is a convex combination of the nursing homes. For example, N4 could be projected onto the best practice frontier at point *X*. The performance measure is the linear distance from the frontier expressed as a percent such as 90, 84% and so on. To be rated 100%, the nursing homes must be on the best practice surface.

In this simple two-dimensional example, the nursing homes with the greatest improvements in functional status were identified. DEA is capable of assessing quality with dozens of indicators in *n*-dimensional space. For a complete mathematical explanation, see Chap. 1 in this book.

A model that could be used to introduce an outcome-quality frontier is called the additive model (described in Chap. 1). The additive to be used is based on the idea of subtracting the functional status of the patient at the outset from the status attained after the care process. The resulting measure is expressed as a change in functional status based on the (functional status achieved during quarter 3) minus (the functional status at quarter 1). The numerical differences from Q_1 to Q_3 can be interpreted as improvements, deteriorations, or no change in outcomes.

The optimal value $w_{o,}^{*}$ is a rating that measures the distance that a particular nursing home being rated lies from the frontier. A separate linear programming model is run for each nursing home (or unit) whose outcomes are to be assessed. The additive model is shown below:

$$
\begin{aligned}
&\text{(additive model)}\\
&\max_{u,v,uo} \; w_o = u^{\mathrm{T}}\mathrm{Y_o} - v^{\mathrm{T}}\mathrm{X_o} + \mathrm{u}_o\\
&\text{s.t.} u^{\mathrm{T}}\mathrm{Y} - v^{\mathrm{T}}\mathrm{X} + \mathrm{u}_o\vec{\mathbf{1}} \le 0\\
&-u^{\mathrm{T}} \le -\vec{\mathbf{1}}\\
&-v^{\mathrm{T}} \prec -\vec{\mathbf{1}}
\end{aligned}
$$

Here the vector Y represents the observed functional status variables at Q_3 and X represents the initial functional status variables observed at Q_1. The additive model subtracts the functional status variable at Q_3 from those at Q_1. The variables, u^{T} and v^{T} are the weights assigned by the linear program so $u^{\mathrm{T}}\mathrm{Y}$ is the weighted functional status at Q_3 and $v^{\mathrm{T}}\mathrm{X}$ is the weighted functional status at Q_1. The model is constrained so that every nursing home is included in the

optimization such that the range of scores will be between 0 and 100%. The u^T and v^T are constrained to be nonnegative.

For the problem of developing a frontier measure of improvements or deterioration in functional status, the additive has number of advantages (see Cooper et al. 2000b). The model does not focus on a proportional reduction of inputs or an augmentation of outputs. It offers a global measure of a distance from a frontier, by giving an "equal focus on functional status before and after by maximizing the functional improvement between time periods (characterized as the difference between Q_1 and Q_3). Another advantage of the additive model is that it is translation invariant which means we can add a vector to the inputs and outputs and though we get a new dataset, the estimates of best practice and the outcome measures will be the same. This model is applied to a nursing home dataset to find a DEA outcome frontier and a DEA decision-making efficiency frontier.

16.5.1 A Field Test: Combining Outcome Frontiers and Efficiency Frontiers

To illustrate the ideas discussed above, the following DEA model will be used to find a quality-outcome frontier for 476 nursing homes in Massachusetts (the USA). We developed outcome measures for the nursing homes from the Management Minutes Questionnaire (MMQ). This is the case-mix reimbursement tool used in several states in the USA and in particular is used in the state of Massachusetts to pay nursing homes for the services they provide Medicaid residents. The MMQ collects information on the level of assistance that nursing home residents need from staff members to carry out activities of daily living such as dressing, eating, and moving about. Fries (1990) explains that the MMQ index is constructed for each resident based on a spectrum of resident characteristics, each with a specified weight. Values are supposed to correspond to actual nursing times so the total should correspond to total staffing needs for the resident. Weights are derived from expert opinion rather than statistical analysis, and total weights are adjusted with time values added for each of the items measured.

Changes in overall resident functioning (determined by measuring the change in MMQ scores over two quarters) were used as a proxy for quality of care. These variables depict the direction of functional status change (improvement, maintenance or decline) experienced by the residents during the last 6 months. Changes in a positive or static direction (improvement or maintenance) will be used as proxies for high-quality care (controlling for health status) and changes in a negative direction (decline) will be used as a proxy for a decrease in the quality of care.

In each of the nursing homes, the residents' functional status was evaluated. The proportion of residents who were independent in mobility and eating, and

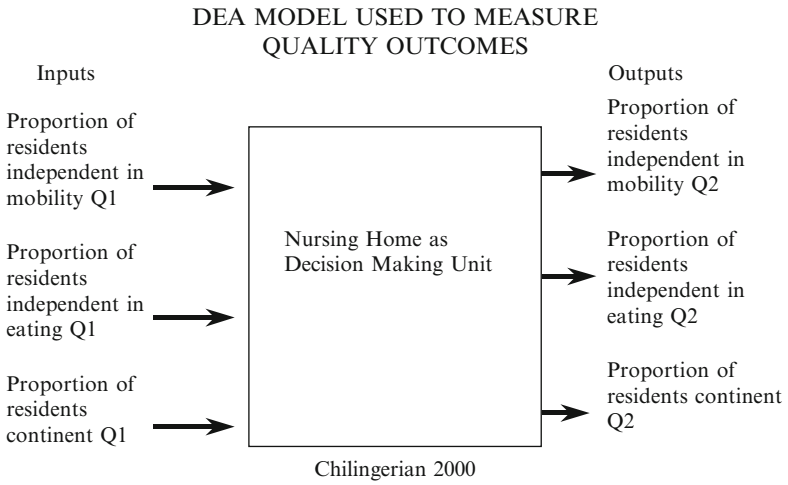


Fig. 16.11 Three-input–three-output model used to estimate quality

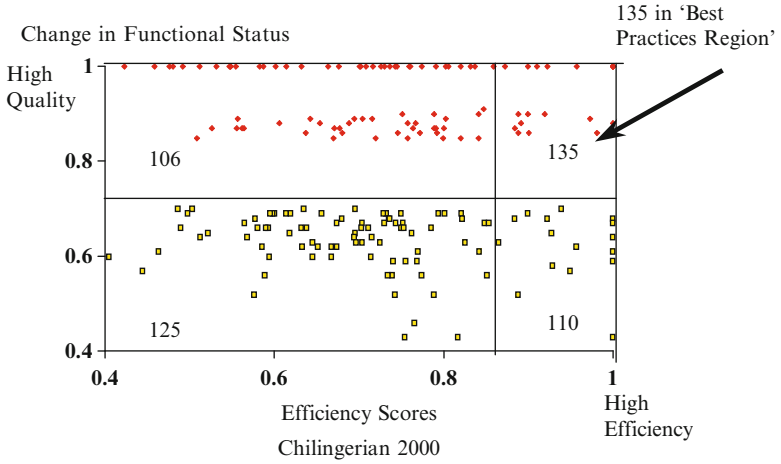


Fig. 16.12 Plot of change in functional status and productive efficiency (source: Chilingerian 2000)

continence were traced for 6 months from quarter 1 to quarter 3. If we consider how a patient's functional status changes over time, whether a nursing home is improving, maintaining, or declining, these changes become an outcome measurement tool. In particular, resident functional improvement was monitored by three activities of daily living: bed mobility, eating, and toilet use. The model used and the variables are displayed in Fig. 16.11 and 16.12.

The DEA model identified 41 or 9% of the nursing homes on the best practice quality frontier. The average quality score was 74%, which means that the functional status outcomes could potentially be improved by 26%.

Figure 16.12 plots the DEA measure of the changes in functional status against a DEA measure of the decision-making efficiency for the 476 nursing homes in the dataset. By partitioning this two-dimensional summary of nursing home performance at the means (81 and 74%), four categories of performers emerge: 135 with high quality outcome and high efficiency; 106 with high quality outcomes but lower efficiency; 110 with low quality outcomes but high efficiency; and 125 with lower quality outcomes and low efficiency. Each of these unique and exhaustive categories can be further analyzed to explore the factors associated with each category of performance. By studying the interaction of these two dimensions of performance, it may be possible to gain new insights into quality and efficiency.

16.6 A Health DEA Application Procedure: Eight Steps

Several studies have suggested analytic procedures for DEA studies (Golany and Roll 1989; Lewin and Minton 1986). Drawing heavily on these papers, this last section highlights how the researcher connects the empirical and conceptual domains with real world problems in health policy and management. The structure of DEA research can be seen as a sequence of the following eight applied research steps:

16.6.1 Step 1: Identification of Interesting Health-Care Problem and Research Objectives

Applied work always begins with a real-life policy and/or management problem. Interesting health policy and management research questions are needed to guide the data to be collected. If we take a close look at all of the DEA studies, many have been illustrations of uses of DEA based on available data. Therefore, the first step is to find questions of practical importance, and never let the available data drive the research. Identify what is new and interesting about the study.

Set challenging research goals – an interesting question is always preferable to a researchable question from an available dataset. The goal should always be to inform both theory and practice. The purpose of a good research question is to get to some answers.

16.6.2 Step 2: Conceptual Model of the Medical Care Production Process

There is no final word on how to frame a health-care production model. Each DEA study claims to capture some part of reality. Hence, the critical question in this step is – Does the production model make sense? Most likely, the problem has been studied before. So an obvious first step is to ask – How has it been modeled in previous work? What aspects were missing? There should always be a justification of the inputs and outputs selected. This can be based on the literature, prior knowledge, and/or expert knowledge (see O'Neill 1998; Chilingerian and Sherman 1997). The use of clinical experts is critical if the application is to become useful for practice.

A related question is choice of the DEA models. Because of strong intuitive appeal, the CCR (clinical and scale efficiency) and BCC (pure clinical efficiency) models have been used in most health-care studies. The multiplicative model (for Cobb–Douglas production functions) and the additive model hardly appear in the health-care literature. Researchers should consider the underlying production technology and take a fresh look at the choice of DEA models. If the justification for the final choice of a DEA model is sketchy, consider running more than one.

16.6.3 Step 3: Conceptual Map of Factors Influencing Care Production

Step three identifies the set of variables and some empirical measures based on several important questions. What is the theory of clinical production or successful performance? Do medical practices vary because some patients are sicker, poorer, or socially excluded? What explains best practices? For this step, researcher should identify the environmental and other factors out of the control of the managers, organizational design and managerial factors, provider and patient characteristics and case-mix variables (see Chilingerian and Glavin 1994). The goal is to build a conceptual map that identifies some of the obstacles, or explanatory variables associated with best practices from the literature and expert knowledge. If the theory is weak, try some simple maps or frames that raise theoretical issues.

16.6.4 Step 4: Selection of Factors

Now the researcher is ready to search for databases or collect the data. There is always some difficulty obtaining the variables for the study files: inputs, outputs, controls, explanatory variables, and the like. Sometimes physician, hospital, medical association, and insurance databases must be merged into one study file to link DEA with other variables.

16.6.5 Step 5: Analyze Factors Using Statistical Methods

DEA assumes that a model is assessing the efficiency of "comparable units," not product differences. Before running an efficiency analysis, if there is reason to believe that outputs are heterogeneous, it is recommended that peer groups be developed (Golany and Roll 1989). In health care, a variety of peer groups could be developed based on medical subspecialty (orthopedic surgeons versus cardiac surgeons), diagnostic complexity, and other product differences.

The mathematics of DEA assumes that there is an isotonic (order preserving) relation between inputs and outputs. An increase in an input should not result in a decrease in an output. Inputs should be correlated. Golany and Roll (1989) have argued that running a series of regressions (variable by variable) can help to reduce these problems. If there is multicollinearity among inputs (or among outputs), one remedy is to eliminate one or more inputs or outputs.

Finally, there has rarely been a DEA study that has not had to deal with the problem of zeroes. In any case, there are two crude ways of dealing with problem. One is to throw out all of the DMUs with missing values and reduce the number of DMUs. The other way is to substitute the zero with a very small number such as 0.001. In addition to this handbook, Charnes et al. (1991) offer better ways of dealing with the problem of zeros in the dataset.

16.6.6 Step 6: Run Several DEA Models

Are the results reasonable? If the dataset finds a threefold difference in costs among the units, something other than efficiency is being measured. If you are running different DEA models check for the stability of results.

Sometimes there are many self-referring units on the best practice frontier. One solution to this problem is to impose cone ratio conditions that reflect preferred practice styles (see Chilingerian and Sherman 1997; Ozcan 1998). As a rule of thumb, if the majority of the DMUs are showing up as 100% efficient and they are mostly self-referencing, there are two possible explanations. Perhaps the production technology is so complex that there are many slightly different ways of practicing. On the contrary, the free choice of weights is giving the providers the "benefit of the doubt" by hiding unacceptable practice styles that care for very few patients, and/or utilize a very high quantity of clinical resources.

16.6.7 Step 7: Analyze DEA Scores with Statistical Methods

The next step is to use the DEA results to test hypotheses about inefficiency. Blending DEA with various statistical methods has been all the rage in health-care

studies. This has been a health trend. To convince the reader that the DEA scores are valid, every health-care study must test the hypothesis that there is unaccounted case mix. If the explanatory and control variables have no significant association with the DEA scores, something may be wrong with the production model.

16.6.8 Step 8: Share Results with Practitioners and Write It Up

Traditionally the DEA work in health care focused more on methodology than the issue of usefulness. To advance the field more quickly, research needs to spend time with clinicians and practitioners (who are not our students). Before writing up the results show the models and findings to real clinical managers and listen to their advice about performance effectiveness. This is a suggestion for everyone.

16.7 DEA Health Applications: Do's and Don'ts

Given the 20-year history we have with DEA health applications, there are several do's and don'ts. Based on our experience on reviewing and reading DEA papers – three are discussed.

16.7.1 Almost Never Include Physicians As a Labor Input

One of the concerns in DEA hospital studies is including physicians as an input in health care (Burgess et al. 1998; Chilingerian and Sherman 1990). In many health-care systems such as those of the USA, Belgium, Switzerland, there are academic medical centers that have salaried physicians reporting to a medical director in the hospital; there are also community general hospitals that have established cooperative, as opposed to hierarchical relations with physicians. Though physicians are granted "admitting privileges" and can use the hospital as a workshop to care for patients, some physicians may not be very active in admitting patients. Physicians can enjoy these privileges at several hospitals (Burns et al. 1994). Including a variable for the quantity of FTE physicians is legitimate input if they are a salaried such as in academic medical centers. However, in many countries community hospitals do not employ physicians so they should not be included as an input.

16.7.2 Use Caution When Modeling Intermediate and Final Hospital Outputs

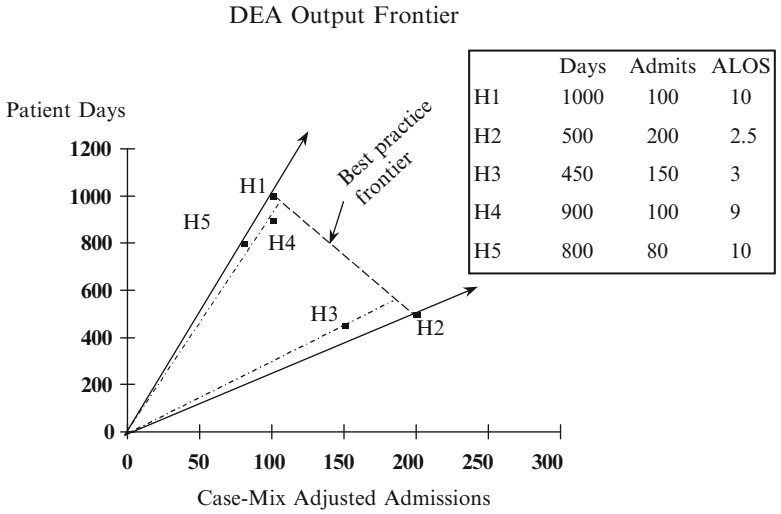
Although modeling hospitals with DEA can be complex, the DEA literature offers some exemplars for research purposes. While none of these papers are perfect, the models are reasonable given the available data. For example:

- Nunamaker (1983) used three outputs: age-adjusted days, routine days, maternity days.
- Sherman (1984) used three outputs: age-adjusted patient days, and nurses and interns trained as outputs.
- Banker et al. (1986) used three outputs: age-adjusted patient days
- Sexton et al. (1989) used six categories of workload-weighted units (for example, medical workload-weighted units (WWU), psychiatric WWU, surgical WWU, OPD WWU, etc.)
- Chilingerian and Sherman (1990) used two outputs: high severity cases of DRG 127 with satisfactory outcomes and low severity cases of DRG 127 with satisfactory outcomes.

Some of these papers focus on "inputs per patient day," while others do on "inputs per discharged case." Note that none of these papers mixed final products, i.e., discharges with intermediate products, i.e., patient days. If both discharges and patient days are the outputs, DEA will obtain results based on a composite of optimal – cost per day/cost per case/cost per visit. DEA will give misleading results.

There have been several papers published in the health-care literature that combine intermediate and final outputs in a single model. While there are arguments on both sides, there are conceptual problems with how DEA would evaluate efficiency using these models. Let us consider the model used by several researchers (see Burgess and Wilson 1996; Ferrier and Valdanis 1996) combining case-mix-adjusted discharges with total patient days. Assuming constant inputs, if a hospital maintains its case-mix-adjusted admissions, while increasing its patient days, its average length of stay (ALOS) is increasing. The DEA model, however, would rate the hospital with higher length of stay (LOS) as more efficient, though these hospitals would, by most managerial definitions, be less efficient because they could have discharged some patients sooner, and *utilized their capacity better* by admitting *more* patients. In a high fixed cost service, longer lengths of stay might suggest poorer quality (more morbidity, nosocomial infections, etc.) Since quality measures are not available, one cannot assume maximum output from inputs with constant quality outcomes.

Now, consider the alternative model with case-mix-adjusted days and teaching outputs, excluding discharges (see Sherman 1994). Assuming constant inputs, if a hospital maintains the number of interns being taught and increases its case-mix-adjusted patient days, the hospital is producing more outputs with fewer inputs – i.e., becoming more efficient. That is, DEA would rate that hospital



	Days	Admits	ALOS
H1	1000	100	10
H2	500	200	2.5
H3	450	150	3
H4	900	100	9
H5	800	80	10

Fig. 16.13 Two-dimensional picture of an output frontier that combines intermediate and final hospital outputs

more efficient. Although the ALOS is not known, the Sherman model is not rating hospitals with longer lengths of stay as better. This type of model works better conceptually and in practice.

To illustrate the point, consider the following data from a policy maker's perspective. In Fig. 16.13 five teaching hospitals, each with *similar inputs* and the same case mix, and the *same total ancillary charges*, *but with variable patient days and case-mix-adjusted admissions*. In the plot of the four hospitals, the DEA output frontier identifies the best practice frontier to be the segment connecting H1 and H2. H4, for example, is less efficient than H1 because it had fewer patient days, even though it discharged the patients sooner and presumably used its available capacity better.

H5 is inefficient and would become more efficient if H5 could move along the ray (represented by the solid lines) and toward H1. The efficiency rating is based on a radial measure of efficiency that maintains the rate of transformation. This measure forces H5 to increase days and admissions, and ignores the need to improve clinical efficiency – e.g., by reducing length of stay and increasing admissions. Do the definitions of efficiency underlying this model make sense?

H4 has the same admissions but fewer patient days; it has the same cost per case, but a higher cost per patient day than H1 and a lower ALOS. From a policy perspective, should not H4 be ranked at least as efficient as H1? DEA suggests that H4 becomes efficient by maintaining its ALOS and increasing days and admissions accordingly.

Now, when H2 is thought of as an extreme point, policy makers might consider H1 as the only frontier point. H2, 3, 4, and 5 all could improve their capacity utilizations (by treating more cases) and improve their clinical efficiency

(by reducing lengths of stay). There are DEA models that define preferred practice regions (or cone ratios) to handle these situations.

Therefore, if the model includes *both* patient days and cases as DEA outputs (as the author's have done), DEA defines two type of efficiency – one based on cost per case, and another based on cost per admission. H1 is rated 100% because it produced more of the variable output – patient days. H2 is rated 100% efficient because it produced more admissions.

Hospitals with the same case mix and admissions, but higher lengths of stay would be rated as 100% efficient along side hospitals with the same case mix, patient days, and lower lengths of stay. This runs counter to what policy makers would want and would confuse everyone. There are two solutions. One is to use the managerial inputs with the clinical outputs (see Fig. 16.4). Another approach is to use a two-part production model: one DEA for practice management with managerial inputs and intermediate outputs, and another DEA model for patient management with clinical inputs and clinical outputs (see Fig. 16.1).

16.7.3 Do Check the Distribution of DEA Scores and Influence of Best Practice Providers on Reference Sets

DEA will yield some approximation of an efficiency score with most datasets. Consequently, every health-care application requires a careful check of the distribution of DEA scores. For example, whenever the range includes efficiency scores below 0.50, or whenever there are more than 50 or 60% of the DMUs on the frontier, there is a strong likelihood that something is wrong. The following common sense test appears to be helpful – Given the health-care context, do these results make sense? Though one can hardly imagine finding a hospital, nursing home, or physician operating at 17% efficiency and surviving another day, it is possible. Low scores always require a plausible explanation, which might be a noncompetitive environment, or the existence of subsidies. Newhouse points out that Zuckerman et al. (1994) found productive inefficiency to account for 14% of total hospital costs (1994). Should policy makers reduce hospital budgets by this amount? Would this force many hospitals out of business? Before anyone takes this findings too seriously, the inefficiencies have to be valid.

In health care, a high proportion of very low DEA scores are likely to be due to unaccounted case mix, heterogeneity of output measures, or returns to scale (Chilingerian 1995). Alternatively, a very high proportion of DMUs on the best practice frontier (40–50%+) could be the result of caring for many different types of patients, using slightly different styles of practice. It is important to examine the number of physicians appearing in only one reference set and to identify the most influential physicians – those physicians who appear in the most reference sets. When a priori information is available on best practices, it is possible to reduce the number of efficiency candidates in any given analysis. For example, a cone ratio DEA model allows for a meaningful upper and lower-bound restriction to be placed

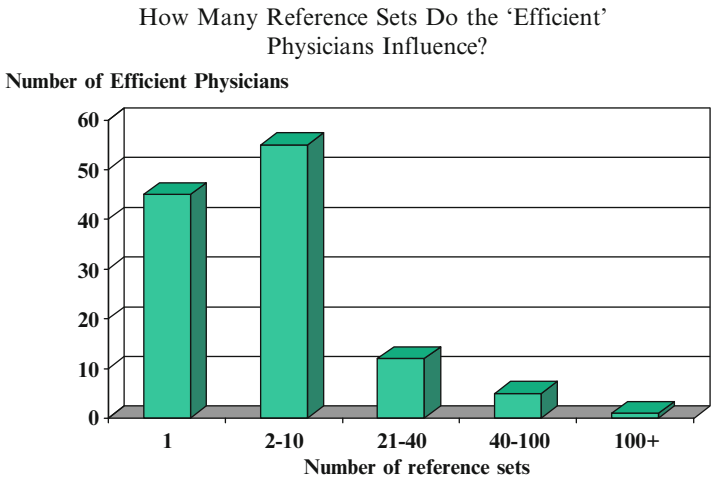


Fig. 16.14 Number of efficient physicians by the number of reference sets they influence

on each input virtual multiplier. The restriction reduces the number of efficient candidates, bringing the DEA measure of technical efficiency closer to a measure of overall efficiency. Imposing cone ratio models on the results will usually reduce the number of self-referencing DMUs on the frontier (see Chilingerian and Sherman 1997; Charnes et al. 1990). Running a second DEA model without the most influential observation will reveal how robust the DEA results.

For example, one DEA study of 326 primary care physicians conducted by Chilingerian and Sherman (1997) found 138 on the best practice frontier. In Fig. 16.14, 45 physicians appear in one reference set and are, therefore, self-referencing physicians. Expert opinion from the clinical director helped to identify a preferred practice region. The medical director's criterion for "best practice" was a primary care physician who (1) performed under budget, (2) utilized less than 369 medical/surgical days per 1,000 members, (3) had referral rates of less than 1.2 per member, and (4) provided at least 1,760 primary care office visits per 1,000 members. A second DEA model would be run setting the minimum and maximum marginal rates of substitution for referral rates, medical surgical hospital days, and primary care visits. The second DEA model reduced the number of efficient physicians from 138 to 85 and found most of the 45 self-referring physicians less efficient.

16.8 A Final Word

This chapter looks at some of the conceptual and methodological challenges associated with measuring and evaluating health-care performance with DEA. While DEA offers many advantages, the critics have raised fundamental questions such as the following: How useful is DEA? What purpose does a DEA study serve? (Newhouse 1994).

There are several purposes for conducting DEA studies. One is to develop better descriptions and analyses about practice patterns and styles. Insights into most optimal caseloads, effects of severity on scale inefficiency, or other sources of inefficiency are helpful to health-care managers.

A second purpose for undertaking a health application with DEA is to identify best practices to create insights and new ideas that explain the successes and failures of clinical providers for policy makers. Examining why some provider firms succeed while others fail and identifying the sources of performance improvement remains an important societal problem, especially if the sources of failure are rooted in the payment and financing systems. Finally, once best practices are identified, health-care mangers need help in finding ways to reduce waste in the utilization of clinical resources to achieve health goals. This strategic objective ranks high on many health managers' agenda.

While the DEA toolbox has a potential to serve these goals, the work has been more illustration than practicable and theory developing. The body of DEA research that has accumulated is substantial. Is there anyway to order and account for the panorama of information? Is there an acceptable general DEA model or subset of models for health and hospital studies? Have DEA applications improved the delivery of care for patients? At this time, the answer to all of these questions is "No." There appears to be neither a taxonomy nor a general single model capable of handling all of the issues that critics can raise. Most distressing is the fact that very few DEA applications in health care have documented any significant improvements to quality, efficiency, or access.

Since the early 1990s, the pace of DEA research has been rapid. Has the information been mined? The problems inherent in deepening understanding of health-care operations and performance today are more intricate than those encountered when DEA research began "exploring and testing" DEA applications to understand health-care performance problems.

The main problem has been the lack of rigor in developing models. Simple measurement errors are less of a problem than selecting patently different input and output variables. To prevent the critics from challenging the models and to help advance policy and management, models should be based on good theoretical ideas. Otherwise, there is a stalemate in the theoretical development of health applications.

This chapter has uncovered research and managerial opportunities and challenges and noted that there is a great deal of work that remains to be done. Health care has a goldmine of clinical information and medical records that document cost, quality, and access. On the one hand, every single diagnosis and illness could be studied with DEA at the physician/patient level. On the other hand, in a single general hospital, there are the more than 5,000 distinct products and services – no single DEA model will ever be able to analyze all of those activities.

In 1996, Seiford asked the experts to nominate "novel" DEA application. Four health-care papers were included among the handful of innovative studies. Every one of these DEA studies focused on individual physicians as DMUs.

This remains a very promising area, since physicians are not only an important provider of care but also the principal decision-makers and entrepreneurs for the care programs and clinical DMUs.

A few years ago Scott (1979) proposed a multi-output-multi-input measure of clinical efficiency that divided the amount of improvement for each type of patient by the cost of each hospitalization (input). Measuring the change in severity of illness from admission to discharge would make a DEA efficiency measure clinically relevant by capturing a more complete picture of a physician's clinical performance relative to the resources used.

Finally, Farrell (1957) has argued that activity analysis can be used to evaluate the productive efficiency of various economic levels ranging from small workshops to an entire industry. Is DEA the right tool to motivate a general model of performance (quality, access, and productive efficiency) in health care? Clearly, DEA is proving to be an effective approach to evaluate individual efficiency as well as organizational efficiency; consequently, DEA could be used to evaluate quality and efficiency at all four levels of the health are industry:

1. The individual patient's experience
2. Individual physician level
3. The department or organizational level
4. The entire industry level

As information systems become more integrated and more available, and as DEA evolves from mere illustration to real health services research, future studies could attempt to connect individual patient changes in health status to physician efficiency, to department efficiency, to overall hospital efficiency, and then connect hospital efficiency to the efficiency of the entire hospital industry.

Acknowledgments We are very grateful to Dianne Chilingerian and W.W. Cooper for their encouragement and thoughtful comments. Part of the material for this chapter was adapted from Chilingerian, J. A., 1995, Evaluating physician efficiency in hospitals: A multivariate analysis of best practices, European Journal of Operational Research 80, 548–574; Chilingerian, J.A. 2000, Evaluating quality outcomes against best practice: a new frontier. The Quality Imperative: Measurement and Management of Quality. Imperial College Press, London, England; and Chilingerian, J. A., and H. David Sherman, 1997, DEA and primary care physician report cards: Deriving preferred practice cones from managed care service concepts and operating strategies, Annals of Operations Research 73, 35–66.

References

Anderson GF, Reinhardt UE, Hussey PS, Petrosyan V. It's the prices, stupid: why the United States is so different from other countries. Health Aff. 2003;22(3):89–105.

Banker RD. Estimating most productive scale size using data envelopment analysis. Eur J Oper Res. 1984;17:35–44.

Banker RD, Conrad R, Strauss R. A comparative application of data envelopment analysis and translog methods: an illustrative study of hospital production. Manage Sci. 1986;32(1):30–44.

Banker R, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci. 1984;30:1078–92.
Banker RD, Morey RC. Efficiency analysis for exogenously fixed inputs and outputs. Oper Res. 1986a;34(4):513–21.
Banker RD, Morey RC. The use of categorical variables in data envelopment analysis. Manage Sci. 1986b;32(12):1613–27.
Bristol Royal Infirmary Inquiry Final Report, The Report of the public inquiry into the children's heart surgery at the Bristol Royal Infirmary, Learning from Bristol. Presented to Parliament by the Secretary of State for Health by command of her Majesty; 2002; London, England: Crown.
Bryce CL, Engberg JB, Wholey DR. Comparing the agreement among alternative models in evaluating HMO efficiency. Health Serv Res. 2000;35:509–28.
Burgess JF, Wilson PW. Hospital ownership and technical efficiency. Manage Sci. 1996;42 (1):110–23.
Burgess J, James F, Wilson PW. Variation in inefficiency among U.S. hospitals. INFOR. 1998;36:84–102.
Burns LR, Chilingerian JA, Wholey DR. The effect of physician practice organization on efficient utilization of hospital resources. Health Serv Res. 1994;29(5):583–603.
Caves RE. Industrial efficiency in six nations. Cambridge: MIT Press; 1992.
Charnes A, Cooper WW, Henderson A. An introduction to linear programming. New York: Wiley; 1953.
Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision-making units. Eur J Oper Res. 1978;3:429–44.
Charnes A, Cooper WW, Thrall R. A structure for classifying and characterizing efficiencies in data envelopment analysis. J Prod Anal. 1991;2:197–237.
Charnes A, Cooper WW, Lewin AY, Seiford LM. Basic DEA models, data envelopment analysis: theory, methodology, and application. Boston: Kluwer Academic Publishers; 1994.
Charnes A, Cooper WW, Sun DB, Huang ZM. Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks. J Econom. 1990;46:73–91.
Chilingerian JA. Investigating non-medical factors associated with the technical efficiency of physicians in the provision of hospital services: a pilot study. Best Paper Proceedings, Annual Meeting of the Academy of Management; 1989; Washington, D.C
Chilingerian JA. Evaluating physician efficiency in hospitals: a multivariate analysis of best practices. Eur J Oper Res. 1995;80:548–74.
Chilingerian JA, Glavin M. Temporary firms in community hospitals: elements of a managerial theory of clinical efficiency. Med Care Res Rev. 1994;51(3):289–335.
Chilingerian JA, Sherman HD. Managing physician efficiency and effectiveness in providing hospital services. Health Serv Manage Res. 1990;3(1):3–15.
Chilingerian JA, David Sherman H. Benchmarking physician practice patterns with DEA: a multi-stage approach for cost containment. Ann Oper Res. 1996;67:83–116.
Chilingerian JA, David Sherman H. DEA and primary care physician report cards: deriving preferred practice cones from managed care service concepts and operating strategies. Ann Oper Res. 1997;73:35–66.
Chilingerian JA. Evaluating quality outcomes against best practice: a new frontier. The quality imperative: measurement and management of quality. London: Imperial College Press; 2000.
Chilingerian JA, Glavin M, Bhalotra S. Using DEA to profile cardiac surgeon efficiency. Draft of Technical Report to AHRQ; 2002; Heller School, Brandeis University.
Cooper WW, Park KS, Pastor JT. Marginal rates and elasticities of substitution with additive models in DEA. J Prod Anal. 2000a;13:105–23.
Cooper WW, Seiford LM, Tone K. Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software. Boston: Kluwer Academic Publishers; 2000b.
Fair R. A theory of extramarital affairs. J Polit Econ. 1978;86:45–61.

Fare R, Grosskopf S, Lindgren B, Roos P. Productivity developments in Swedish pharmacies: a malmquist output index approach, data envelopment analysis: theory, methodology, and applications. Boston: Kluwer Academic Publishers; 1994.
Farrell MJ. The measurement of productive efficiency. J R Stat Soc Ser A. 1957;120:253–81.
Ferrier GD, Valdanis V. Rural hospital performance and its correlates. J Prod Anal. 1996;7:63–80.
Fetter RB, Freeman J. Diagnostic related groups: product line management with in hospitals. Acad Manage Rev. 1986;11(1):41–54.
Finkler MD, Wirtschafter DD. Cost-effectiveness and data envelopment analysis. Health Care Manage Rev. 1993;18(3):81–9.
Fizel JL, Nunnikhoven TS. Technical efficiency of nursing home chains. Appl Econ. 1993;25:49–55.
Fries BE. Comparing case-mix systems for nursing home payment. Health Care Financ Rev. 1990;11:103–14.
Georgopoulos BS. Organizational problem solving effectiveness: a comparative study of hospital emergency services. San Francisco: Jossey-Bass Publishers; 1986.
Golany B, Roll Y. An application procedure for DEA. Omega. 1989;17(3):237–50.
Greene WH. Econometric analysis. 2nd ed. New York: Macmillan Publishing Company; 1983.
Harris JE. The internal organization of hospitals: some economic implications. Bell J Econ. 1977;8:467–82.
Harris JE. Regulation and internal control in hospitals. Bull N Y Acad Med. 1979;55(1):88–103.
Hao S, Pegels CC. Evaluating relative efficiencies of veteran's affairs medical centers using data envelopment, ratio, and multiple regression analysis. J Med Syst. 1994;18:55–67.
Hollingsworth B, Dawson PJ, Maniadakis N. Efficiency measurement of health care: a review of non-parametric methods and applications. Health Care Manage Sci. 1999;2(3):161–72.
Kerr C, Glass JC, McCallion GM, McKillop DG. Best-practice measures of resource utilization for hospitals: a useful complement in performance assessment. Public Adm. 1999;77(3):639–50.
Kooreman P. Nursing home care in The Netherlands: a non[parametric efficiency analysis. J Health Econ. 1994;13:301–16.
Lapetina EM, Armstrong EM. Preventing errors in the outpatient setting. Health Aff. 2002;21 (4):26–39.
Leibenstein H, Maital S. Empirical estimation and partitioning of x-inefficiency: a data envelopment approach. Am J Econ. 1992;82(2):428–33.
Lewin AY, Minton JW. Determining organizational effectiveness: another look and an agenda for research. Manage Sci. 1986;32(5):514–38.
Lynch J, Ozcan Y. Hospital closure: an efficiency analysis. Hosp Health Serv Adm. 1994;39 (2):205–12.
Luke RD, Ozcan YA. Local markets and systems: hospital consolidations in metropolitan areas. Health Serv Res. 1995;30(4):555–76.
Maddala GS. Limited dependent and qualitative variables in econometrics. New York: Cambridge University Press; 1983.
Morey D, Ozcan Y. Med Care. 1995;5:531–52.
Newhouse JP. Frontier estimation: how useful a tool for health economics? J Health Econ. 1994;13:317–22.
Newhouse J. Why is there a quality chasm? Health Aff. 2002;21(4):13–25.
Neuman BR, Suver JD, Zelman WN. Financial management: concepts and applications for health care providers. Maryland: National Health Publishing; 1988.
Nunamaker T. Measuring routine nursing service efficiency: a comparison of cost per day and data envelopment analysis models. Health Serv Res. 1983;18(2 Pt 1):183–205.
Nyman JA, Bricker DL. Profit incentives and technical efficiency in the production of nursing home care. Rev Econ Stat. 1989;56:586–94.
Nyman JA, Bricker DL, Link D. Technical efficiency in nursing homes. Med Care. 1990;28:541–51.

Ozcan Y. Physician benchmarking: measuring variation in practice behavior in treatment of otitis media. Health Care Manage Sci. 1998;1:5–18.
Ozcan Y, Lynch J. Rural hospital closures: an inquiry into efficiency. Adv Health Econ Health Serv Res. 1992;13:205–24.
Ozcan YA, Begun JW, McKinney MM. Benchmarking organ procurement organizations: a national study. Health Serv Res. 1999;34:855.
Ozcan YA, Luke RD. A national study of the efficiency of hospitals in urban markets. Health Serv Res. 1992;27:719–39.
Ozgen H, Ozcan YA. A national study of efficiency for dialysis centers: an examination of market competition and facility characteristics for production of multiple dialysis outputs. Health Serv Res. 2002;37:711–22.
Perez ED. Regional variations in VAMC's operating efficiency. J Med Syst. 1992;16(5):207–13.
Puig-Junoy J. Technical efficiency in the clinical management of critically ill patients. Health Econ. 1998;7:263–77.
Rosko MD. Measuring technical efficiency in health care organizations. J Med Syst. 1990;14 (5):307–22.
Rosko MD, Chilingerian JA, Zinn JS, Aaronson WE. The effects of ownership, operating environment, and strategic choices on nursing home efficiency. Med Care. 1995;33 (10):1001–21.
Scott RW. Measuring outputs in hospitals, measuring and interpreting productivity. Nat Res Counc. 1979;255–75.
Seiford L. Data envelopment analysis: the satt6e of the art. J Prod Anal. 1996;7(2/3):99–138.
Sexton T, Leiken A, Nolan A, Liss S, Hogan A, Silkman R. Evaluating managerial efficiency of veterans administration medical centers using data envelopment analysis. Med Care. 1989;27 (12):1175–88.
Sherman HD. Hospital efficiency measurement and evaluation. Med Care. 1984;22(10):922–8.
Shortell S, Kaluzny A. Health care management. 4th ed. New York: Delmar Publishers Inc.; 2000.
Teboul J. Le temps des services: une nouvelle approche de management. Paris: Editions d'Organisation; 2002.
Witte A. Estimating an economic model of crime with individual data. Q J Econ. 1980;94:57–84.
Zhu J. Further discussion on linear production functions and DEA. Eur J Oper Res. 2000;127:611–8.
Zuckerman S, Hadley J, Iezzoni L. Measuring hospital efficiency with frontier cost functions. J Health Econ. 1994;13:255–80.

Index

W.W. Cooper et al. (eds.), *Handbook on Data Envelopment Analysis*,
International Series in Operations Research & Management Science 164,
DOI 10.1007/978-1-4419-6151-8,