International Series in
Operations Research & Management Science

Juan Aparicio
C. A. Knox Lovell
Jesus T. Pastor *Editors*

Advances in Efficiency and Productivity

Springer

International Series in Operations Research & Management Science

Volume 249

More information about this series at http://www.springer.com/series/6161

Juan Aparicio · C. A. Knox Lovell
Jesus T. Pastor
Editors

Advances in Efficiency and Productivity

Springer

Editors
Juan Aparicio
Center of Operations Research
Miguel Hernández University of Elche
Elche, Alicante
Spain

Jesus T. Pastor
Center of Operations Research
Miguel Hernández University of Elche
Elche, Alicante
Spain

C. A. Knox Lovell
School of Economics, CEPA
University of Queensland
Brisbane, QLD
Australia

ISSN 0884-8289 ISSN 2214-7934 (electronic)
International Series in Operations Research & Management Science
ISBN 978-3-319-48459-4 ISBN 978-3-319-48461-7 (eBook)
DOI 10.1007/978-3-319-48461-7

Library of Congress Control Number: 2016955078

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Part I
Background

Chapter 1
Editors' Introduction

Juan Aparicio, C. A. Knox Lovell and Jesus T. Pastor

Abstract We begin by providing some historical background behind this book. We continue by discussing the significance of the three operative words in the title of the book—advances, efficiency and productivity. We then briefly summarize the chapters in the book, which divide into advances in the analytical foundations of efficiency and productivity, and advances in empirical applications of efficiency and productivity.

Keywords Efficiency · Productivity

1.1 Background

The Santander Chair of Efficiency and Productivity was created at the Miguel Hernandez University (UMH) of Elche, Spain, at the end of year 2014. Its aim is to promote specific research activities among the international academic community. This Research Chair was ascribed to The UMH Institute Center of Operations Research (CIO). The funding of the Chair by Grupo Santander constitutes one more example of the generosity and the vision of this organization, which supports a network of over 1400 Ibero-American universities, covering 23 countries and over 19 million students and academicians. As Director of the Chair was appointed Prof. Knox Lovell, Honorary Professor of Economics at the University of Queensland,

J. Aparicio · J.T. Pastor (✉)
Center of Operations Research (CIO), University Miguel Hernández of Elche, Elche, Spain
e-mail: jtpastor@umh.es

J. Aparicio
e-mail: j.aparicio@umh.es

C.A. Knox Lovell
Centre for Efficiency and Productivity Analysis, University of Queensland, Brisbane, Australia
e-mail: k.lovell@uq.edu.au

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_1

Australia. Further, the Advisory Board is integrated by four other members, two of them on behalf of Grupo Santander, Mr. José Marìa García de los Ríos and Mr. Joaquín Manuel Molina, and the other two on behalf of the UMH, Ph.D. Juan Aparicio, appointed as Co-Director, and Ph.D. Lidia Ortiz, the Secretary of the Chair. During 2015 and 2016 the Chair has organized eight Efficiency/Productivity Seminars for starting new programs with researchers interested in a variety of topics such as Education, Municipalities, Financial Risks, Regional Cohesion, Meta-heuristics, Renewable Energy Production, Food Industry and Endogeneity. During 2015 an International Workshop on Efficiency and Productivity was organized. The Workshop contributions of fifteen relevant researchers/research groups made it possible to conceive this book, entitled "Advances in Efficiency and Productivity", with the inestimable support of Professor Joe Zhu, the Associate Series Editor for the Springer International Series in Operations Research and Management Sciences.

1.2 Advances in Efficiency and Productivity

The title of this book is generic, but the two substantive economic components, efficiency and productivity, are of great importance to any business or economy, and the first word in the title is the theme of the book and of the 2015 International Workshop on Efficiency and Productivity that spawned it. The presentations at the workshop, and the chapters in this book, truly do advance our understanding of efficiency and productivity.

The theoretical definition of efficiency involves a comparison of observed inputs (or resources) and outputs (or products) with what is optimal. Thus technical efficiency is the ratio of observed to minimum input use, given the level of outputs produced, or as the ratio of observed to maximum output production, given the level of input use, or some combination of the two. Each type of technical efficiency is conditional on the technology in place, and each is independent of prices. In contrast, economic efficiency is price-dependent, and typically is defined as the ratio of observed to minimum cost, given outputs produced and input prices paid, or as the ratio of observed to maximum revenue, given inputs used and output prices received. There are variants of these definitions based on different measures of value, such as the ratio of observed to maximum profit, but all take the form of a comparison of observed values with optimal values, conditional on the technology in place. The principal challenge in efficiency measurement is the definition of minimum, maximum or optimum, each of which is a representation of the unobserved production technology, which must therefore be estimated. The analytical techniques developed in this book provide alternative ways of defining optimum, typically as a (technical) production frontier or as an (economic) cost, revenue or profit frontier, and alternative ways of measuring efficiency relative to an appropriate frontier.

The theoretical definition of productivity coincides with the commonsense notion of the ratio of output to input. This definition is straightforward on the rare occasion in which a producer uses a single input to produce a single output. The definition is more complicated otherwise, when multiple outputs in the numerator must be aggregated using weights that reflect their relative importance, and multiple inputs in the denominator must be aggregated in a similar fashion, so that productivity is again the ratio of two scalars, aggregate output and aggregate input. The time path of aggregate output is an output quantity index, and the time path of aggregate input is an input quantity index. Productivity growth is then defined as the rate of growth of the ratio of an output quantity index to an input quantity index or, alternatively, as the rate of growth of an output quantity index less the rate of growth of an input quantity index. Market prices are natural choices for weights in the two indexes, provided they exist and they are not distorted by market power or other phenomena. The principal challenge in productivity measurement occurs when prices do not exist or are unreliable indicators of the relative importance of corresponding quantities. In this case productivity and its rate of growth must be estimated, rather than calculated from observed quantities and prices. The analytical techniques developed in this book provide alternative methods for estimating productivity, and productivity change through time or productivity variation across producers.

The concepts of efficiency and productivity are significant beyond academe, and characterize two important dimensions of the performance of businesses and economies, as evidenced by the following examples.

- The financial performance of businesses depends on the efficiency with which they conduct their operations. Arthur D. Little, a consultancy, has ranked 51 European banks by their cost-income ratio, a traditional measure of operating efficiency in banking, over 2004–2006. It found the ten most efficient banks to have a cost-revenue ratio of 45%, and the ten least efficient banks to have a cost-revenue ratio of 72%. It seems worthwhile to search for the sources of the variation in operating efficiency, if only in an effort to shore up the financial performance of the laggard banks, some of which are very large.
- The relative prosperity of economies depends in large part on their productivity growth. According to the OECD, labor productivity (real GDP per hour worked) has grown faster in Germany than in France and Italy over the period 2007–2012. Not coincidentally, according to the World Bank, GDP per capita also has grown faster in Germany than in France and Italy every year for the past decade. Many of the sources of the productivity gaps are well known, but a decade of experience suggests that they are difficult to rectify.

To be useful, efficiency and productivity must be not just well defined, but capable of measurement using quantitative techniques. Many popular concepts, such as cost-income ratios, labor productivity and GDP per capita, can be calculated directly from company reports and country national accounts. When the data constraint eliminates direct calculation, empirical efficiency and/or productivity

analysis is required. Such analyses typically are based on either a mathematical programming technique known as data envelopment analysis (DEA) or an econometric technique known as stochastic frontier analysis (SFA). Both types of estimation construct best practice frontiers, technical or economic, that bound the data from above or below, and these frontiers provide empirical approximations to the theoretical optima referred to above. Both types of estimation, but especially DEA, are analyzed and employed in this book.

1.3 The Contents of the Book

The contributions serendipitously allocate themselves almost evenly between purely analytical chapters that advance our knowledge of the theory and modeling of production and productivity, and chapters that provide detailed empirical applications of analytical concepts that expand our understanding of the roles of efficiency and productivity in influencing the performance of various sectors of the economy.

1.3.1 Analytical Foundations

The eight explorations into the analytical foundations of efficiency and productivity analysis range broadly across modeling issues, from technological possibilities to modeling the structure of production technology to measuring productivity change to purely computational issues. Four contributions are independent of the way production technology is modeled, and four contributions are expressed within a DEA framework. Three contributions provide brief empirical applications designed to illustrate the theoretical advances.

1.3.1.1 Modeling Advances

In Chap. 2 Pastor, Aparicio, Alcaraz, Vidal and Pastor introduce a novel concept, the "reverse directional distance function," that allows to express any single-valued DEA inefficiency measure as a true directional distance function. As an obvious consequence, any property of a directional distance function is satisfied by the single-valued DEA inefficiency measure. For instance, the typical two-step procedure that transforms a DEA measure into a comprehensive DEA measure, i.e., a measure that projects any inefficient point to the strongly efficient subset of the frontier, can be applied to any directional distance function. In fact, concatenating the standard directional vector $g = (-g_x, g_y)$ of a given directional distance function with an additive function that guarantees that optimal projections for all inefficient units belong to the strongly efficient subset, gives rise to a comprehensive reverse

directional distance function that solves the initial problem. The introduced approach has been designed for any DEA measure. In case we are facing a multiple-valued DEA measure, we need to select first one of the associated single-valued measures and proceed with it. A specific criterion can be considered for making the last mentioned choice, as explained in the chapter. The authors then show how to use a reverse directional distance function to estimate cost, revenue or profit inefficiency and decompose it into its technical and allocative inefficiency components. Again, the decomposition is valid for any DEA inefficiency measure that has been expressed previously as a reverse directional distance function. The principles are general, but their analysis is conducted within a DEA framework.

In Chap. 3 Asmild and Hougaard develop a new method for estimating technological production possibilities, which are typically estimated with frontier techniques, either DEA or SFA. The authors construct a hyper-volume index that compares two feasible production possibility sets, say those of public and private service providers, in a way that overcomes some shortcomings of previous efforts to construct an encompassing pooled frontier from two or more group frontiers. The hyper-volume index compares dominance volumes for two feasible production possibility sets. The authors acknowledge computational and unequal sample size issues. They apply the index to an acquiring company and an apparently better performing company it acquires, in an effort to determine whether the apparent superior performance of the acquired company had superior technology or superior management. Based on the hyper-volume indices for the two companies, the authors conclude that the acquiring company did not acquire a better technology, but did acquire a better managed (i.e., more efficient) company.

In Chap. 4, Aparicio, Borras, Pastor and Zofio, inspired by Debreu's 'dead loss' function, introduce a new distance function they call a loss distance function which, besides providing a traditional way of characterizing the structure of production technology, has the property of generalizing dual results of all the already known distance functions. The authors state monotonicity, curvature and other properties of loss distance functions, and compare them with other distance function representations of production technology. The authors show that, under appropriate normalization conditions, loss distance functions fully characterize the considered technology. They also demonstrate a duality between loss distance functions and Hotelling's profit function, and they show that this duality relationship encompasses all previously known duality results. Finally, under differentiability they derive Hotelling's lemma and show that the optimal value of the Lagrange multiplier associated with a well-behaved normalization condition is the loss distance function at the corresponding point.

In Chap. 5 Grifell-Tatjé and Lovell analyze productivity and price recovery indices, the first of which is based on distance functions and the second on value, revenue or cost, frontiers. They generalize previous results by deriving, and providing economic interpretations for, exact relationships between (i) empirical Fisher quantity, productivity, price and price recovery indices and (ii) theoretical Malmquist quantity and productivity indices and Konüs price and price recovery indices. It is well known that empirical indices approximate their theoretical

counterparts, some more closely than others, but the nature of the approximations has been left unclear. The authors construct quantity mix and price mix functions that make previously known approximate relationships exact. These mix functions involve the allocative efficiency of pairs of quantity vectors and price vectors. Their introduction enables our best theoretical quantity and productivity indices (Malmquist) and empirical price and price recovery indices (Fisher), and our best empirical quantity and productivity indices (Fisher) and theoretical price and price recovery indices (Konüs), to satisfy the product test with the relevant value change.

1.3.1.2 Extensions of DEA

In Chap. 6 Førsund relates technical efficiency measures to productivity measures, using well-behaved aggregator functions for outputs and inputs, and relates both to neoclassical production frontiers satisfying Frisch's *Regular Ultra Passum* law. He then extends the well-behaved but unknown aggregator functions to a Malmquist productivity index, which can be estimated and decomposed using DEA. The author notes, significantly, that the Malmquist productivity index is defined on a benchmark technology satisfying global constant returns to scale, which differs from the best practice technology having the structure of a neoclassical production frontier allowing for a pattern of increasing, constant and decreasing returns to scale as size increases. The difference between the two reflects a size effect intended to capture the productivity impact of producing at a point on the best practice technology other than that which maximizes productivity, and for which the scale elasticity is unity.

Chapters 7 and 8 examine weight restrictions in DEA in different contexts. Weight restrictions are imposed when the endogenously determined DEA multipliers (or weights) are deemed unsatisfactory, for whatever reason.

In Chap. 7 Ramón, Ruiz and Sirvent explore the implications for benchmarking and target setting of imposing weight restrictions on DEA models. At issue is whether such models generate infeasible targets and unattainable benchmarks, and the answer depends on whether the weight restrictions reflect management preferences representing value judgements independent of technology, which allows infeasibility, or engineering features of the technology not otherwise incorporated in the DEA framework, which guarantee feasibility. The authors develop primal (envelopment) and dual (multiplier) DEA models with AR-1 type weight restrictions augmented with additional restrictions to generate closest strongly efficient feasible targets, regardless of the source of the weight restrictions. The additional restrictions are key. The authors illustrate their framework by setting targets in a previously analyzed sample of hospitals.

In Chap. 8 Agrell and Bogetoft examine weight restrictions in a regulatory setting. DEA generates an endogenous weight vector for each production unit, and these vectors can vary among units to an undesirable extent. Weight restrictions limit the variation. The authors modify the standard DEA envelopment program to derive, and provide a regulatory interpretation of, a single endogenous vector of

industry-wide weights common to all firms that maximizes the total payment the industry can claim from the regulator. As the solution to a modified DEA problem, this common weight vector remains endogenous. The authors also show that the modified dual multiplier program seeks the maximum payment to the industry, and that the multipliers are the common endogenous weights. The authors provide an illustration based on a sample of district heating plants in Denmark.

In Chap. 9 Aparicio, Gonzalez, Lopez-Espin and Pastor contrast the usual DEA practice of estimating technical efficiency radially relative to (possibly distant) weakly efficient targets located on the best practice frontier with an alternative practice based on the Principle of Least Action, which seeks the closest strongly efficient targets. While the latter practice is sensible from both theoretical and business perspectives, it is computationally more demanding because it is not based on solving a linear program. The authors recognize such problems as combinatorial NP-hard problems, which they solve using genetic algorithms and heuristics, and they provide several numerical illustrations.

1.3.2 Empirical Applications

The seven applications of production theory range broadly across sectors of the economy. Five contributions are based on individual production units and two are based on aggregates of production units. Four contributions use variants of DEA, and three use variants of SFA.

1.3.2.1 Individual Production Units

In Chap. 10 Bos, van Lamoen and Sanders study what they call innovative firms, firms that use innovation inputs to produce innovation outputs. They use stochastic frontier analysis to estimate a knowledge production frontier, which they use in turn to account for and explain observed heterogeneity of productivity (they call it "innovativity") in an unbalanced firm-level panel data set in the Netherlands. They then decompose variation in innovativity into variation in knowledge production technology and variation in the efficiency with which firms use innovation inputs to produce innovation outputs, which in turn depends on several contextual variables. The authors creatively specify a knowledge output and the knowledge inputs used to produce it, and the contextual variables as well, and they use SFA to estimate the knowledge production frontier. Among their findings, variation in the efficiency with which innovation inputs are used to create innovation output explains most of the observed heterogeneity.

In Chap. 11 Calleja-Blanco and Grifell-Tatjé study the impact of potential coopetition, defined as cooperation among competitors, on the financial performance of European automobile plants. They use return on assets as a measure of financial performance, and they invoke the duPont triangle to decompose the

impact of potential coopetition on ROA into its impact on the profit margin and its impact on asset turnover. Their data set is a large unbalanced panel of production plants in 18 of the EU-28 countries observed over 12 years. They use DEA to implement the decomposition. They find large potential ROA gains from coopetition, and these gains occur through both legs of the duPont triangle. They also find the share occurring through the profit margin leg to be attributable primarily to productivity gains. In light of the potential productivity gains and ROA gains available through coopetition, the authors conclude by pondering the legal status of coopetition.

In Chap. 12 Orea and Wall use SFA to estimate the eco-efficiency of a sample of dairy farms in the Spanish region of Asturias. They adopt the standard definition of eco-efficiency as the ratio of economic value added to environmental pressure, and they model this ratio with a DEA-generated index of the ability to maximize profit subject to endogenously determined non-negative weights on a vector of environmental pressures. However they depart from this approach by specifying and estimating a single-stage heteroskedastic SFA model that estimates the weights, and in which the non-negative inefficiency error component is a function of contextual variables, a vector of farmer attitudes and characteristics. They find substantial variation in eco-efficiency across farms, and non-negligible impacts of farmer attitudes and characteristics on eco-efficiency.

In Chap. 13 Kingyens, Paradi and Tam develop a non-oriented slacks-based DEA bankruptcy prediction model based on information available from annual reports, including financial statements and their accompanying notes, particularly the management discussion and analysis, and the auditor's report. They apply this model to a large panel of US retail apparel firms in a competitive industry characterized by low profit margins and frequent turnover of firms. The authors estimate three DEA models, using a layering technique to peel off the most efficient firms, eventually leaving a subset of firms having a high risk of bankruptcy. The authors calculate type I and type II errors for predictions one, two and three years in advance, and they find that their model that incorporates management decision-making information from annual reports out-predicts a popular bankruptcy prediction model.

In Chap. 14 Almanidis and Sickles combine a stochastic cost frontier analysis model with a mixture hazard model to explore the determinants of survival and failure for a panel of US commercial banks during the 2007–2011 financial crisis. Some of these banks failed, but most survived. The combined model estimates the probability and time to failure conditional on a bank's performance. The authors also calculate predictive accuracy based on type I errors, type II errors and overall classification errors, both in-sample and out-of-sample, to assess their potential to serve as early-warning models. Among their findings, estimated cost efficiency is marginally higher for non-failed banks than for failed banks; type I, type II and overall classification errors are impressively small, even for out-of-sample classifications; and capital adequacy and non-performing loans performed important signaling roles, although Federal Reserve System membership sent a negative signal, suggesting a moral hazard problem.

1.3.2.2 Aggregates of Production Units

In Chap. 15 Arocena, Gómez-Plana and Peña explore the evolution of energy intensity, the ratio of energy consumption to output, in nine Spanish manufacturing industries between 1999 and 2007. Change in energy intensity is initially decomposed into an intensity effect accounting for change in energy intensity within industries, and a structural effect accounting for change in industry shares in aggregate output. The intensity effect is decomposed, using frontier techniques, into five sources, including the traditional efficiency and technical change effects and a regional effect. The authors find that improvements in technology, adjustments to the input mix away from energy consumption, and improvements in energy efficiency are the primary sources of reduced energy intensity in Spanish manufacturing, and that trends in the regional distribution of production have tended to increase energy intensity.

In Chap. 16 Fried and Tauer examine the productivity of US farmers as they age, the policy issue being couched as the hypothesis that older farmers are less productive than younger farmers. They test the hypothesis using DEA to construct a Malmquist productivity index, with a very interesting twist. Productivity indices typically are estimated using panel data consisting of a number of production units observed over multiple time periods. The authors use a single cross-section consisting of state-level data observed over multiple age cohorts. The mathematics is unchanged, but the economic interpretation shifts from productivity change through time to productivity variation across age cohorts. The main finding is that, after the 35–44 age cohort, productivity tends to increase with age, contrary to the worry expressed in the title of the chapter. In addition to experience, the authors point to changes in farming technology that have made farming a less physical occupation than it once was as drivers of their finding.

Acknowledgements We gratefully acknowledge the support of Grupo Santander for the organization of the 2015 Workshop on Efficiency and Productivity, as well as to the members of the Advisory Board of the Chair who were engaged in its design and development. We would also like to express our gratitude to Springer Verlag and specifically to Professor Joe Zhu who encouraged us to publish this book. Finally, we thank our authors for their participation at the Workshop and their contributions to this book.

Part II
Analytical Foundations

Chapter 2
The Reverse Directional Distance Function

Jesus T. Pastor, Juan Aparicio, Javier Alcaraz, Fernando Vidal and Diego Pastor

Abstract The aim of any Data Envelopment Analysis (DEA) inefficiency model is to calculate the efficient projection of each unit belonging to a certain finite sample. The reverse directional distance function (*RDDF*) is a new tool developed in this chapter that allows us to express any known DEA inefficiency model as a directional distance function (*DDF*). Hence, given a certain DEA inefficiency model, its *RDDF* is a specific *DDF* that truly reproduces the functioning of the considered DEA model. Automatically, all the interesting properties that apply to any *DDF* are directly transferable to the considered DEA model through its *RDDF*. Hence, the *RDDF* enlarges the set of properties exhibited by any DEA model. For instance, given any DEA inefficiency model, its economic inefficiency—in any of its three possible versions—, can be easily defined and decomposed as the sum of technical inefficiency and allocative inefficiency thanks to the *RDDF*. We further propose to transform any non-strong *DDF* into a strong *DDF*, i.e., into a *DDF* that projects all the units onto the strongly efficient frontier. This constitutes another indication of the transference capacity of the *RDDF*, because its strong version constitutes in itself a strong version of the original DEA model considered. We further propose to search for alternative projections so as to minimize profit inefficiency, and add an appendix showing how to search for multiple optimal solutions in additive-type models.

Keywords Data envelopment analysis · Reverse directional distance function · Economic inefficiency decomposition · Strong DEA models

J.T. Pastor · J. Aparicio (✉) · J. Alcaraz
Center of Operations Research (CIO), Miguel Hernandez University of Elche (UMH), 03202 Elche (Alicante), Spain
e-mail: j.aparicio@umh.es

F. Vidal
Economics, Sociology and Agricultural Policy, Miguel Hernandez University of Elche (UMH), 03212 Orihuela (Alicante), Spain

D. Pastor
Physical and Sports Education Miguel Hernandez University of Elche (UMH), 03202 Elche (Alicante), Spain

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_2

2.1 Introduction

This "Introduction" comprises a description of the subsequent sections as well as a revision of the previous related literature. In Sect. 2.2, we start showing how to transform any DEA inefficiency model into an equivalent *DDF*, called *RDDF*. As mentioned before, the advantages of the *RDDF* is that it has, as a *DDF*, interesting properties, which are inherited by the DEA model where it comes from.

Any time we consider a strong DEA model, i.e., a model whose projections belong to the strongly efficient frontier, the corresponding *RDDF* is also a strong *DDF*. Otherwise, we propose in Sect. 2.3 a method for transforming a weak *DDF*, that is, a *DDF* whose set of projections does not belong to the strong frontier, into a strong *DDF*. Our method identifies the set of strongly efficient projections that defines the new strong *DDF*. In particular, if we want to transform a weak DEA inefficiency measure into a strong one, we always have the option of working with its associated *RDDF*. Consequently, this methodology solves the general problem of transforming any weak DEA inefficiency measure into a strong DEA inefficiency measure. We are only aware of a previous paper that transforms two specific weak DEA efficiency measures into strong DEA efficiency measures (see Asmild and Pastor 2010). In order to derive a comprehensive inefficiency measure associated to the generated strong *DDF*, all we have to do is to consider, at each unit being rated, a strong directional vector that is comparable with the original directional vector associated to the weak *DDF*. We close Sect. 2.3 by making a simple proposal that basically pursues the notion that the two directional vectors at each point have the same Euclidian length.

Section 2.4 extends the findings of Sect. 2.3 to any DEA inefficiency model, *M*, through the corresponding *RDDF*. If it happens that its *RDDF* is a weak *DDF*, the tools developed in Sect. 2.3 are directly applied in order to generate a strong RDDF. This strong *RDDF* is associated to *M* and offers a comprehensive inefficiency measure for it. Consequently, we have solved the problem of associating to any DEA non-comprehensive inefficiency model a *DDF* comprehensive inefficiency model. Moreover, we apply this result to generate, for the first time, comprehensive radial inefficiency models.

Overall inefficiency measurement and decomposition are important for firms facing a world of changing prices since the resultant loss has implications on managers' decision making. In standard microeconomic theory, the economic behavior of a DMU (Decision Making Unit) is usually characterized by cost minimization, revenue maximization, or profit maximization. In particular, if profit maximization is assumed, the DMU faces exogenously determined market output and input prices, and we may assume that the objective of each DMU is to choose the output combination that yields the maximum profit efficiency. In this sense, profit efficiency indicates how close the actual profit of the evaluated DMU approaches the maximum feasible profit. Additionally, in the Farrell (1957) tradition, overall efficiency has usually been decomposed into the product of two

components, technical efficiency and allocative efficiency, as a way to understand what needs to be done to enhance the performance of the assessed unit.

Chronologically, the empirical estimation of technologies from a dataset began in the area of economics with the application of regression analysis and Ordinary Least Squares (OLS) to estimate a parametrically specified 'average' production function (see, e.g., Cobb and Douglas 1928). Later, Farrell (1957) was the first in showing, for a single output and multiple inputs, how to estimate an isoquant enveloping all the observations. Farrell also showed how to decompose cost efficiency into technical and allocative efficiencies. In his paper one can find the first practical implementation of the Debreu coefficient of resource utilization (Debreu 1951) and the Shephard input distance function (Shephard 1953). Farrell's paper inspired other authors to continue this line of research by either a non-parametric piece-wise linear technology or a parametric function. The first possibility was taken up by Charnes et al. (1978) and Banker et al. (1984) resulting in the development of DEA radial models, closely related to the Shephard distance functions; while the latter approach was taken up at the same time by Aigner et al. (1977), Battesse and Corra (1977) and Meeusen and van den Broeck (1977), subsequently resulting in the development of the stochastic frontier models.

As previously mentioned, the decomposition proposed by Farrell was inspired on the work of Shephard, in the sense that the technical efficiency component is really the inverse of the Shephard input distance function. Indeed, Shephard (1953) also defined an output-oriented distance function and established several dual relationships. Much later, Färe and Primont (1995) developed a dual, but not natural, correspondence between Shephard's distance functions and the profit function. In recent years there has been extensive interest in the duality theory and distance functions as can be easily checked. If one defines an optimization problem with respect to quantities, then a dual problem can be defined with respect to (shadow) prices that has the same value. This approach is of great interest for microeconomics both for understanding the mathematics and for clarifying the economics. Chronologically speaking, Luenberger (1992a, b) and later Chambers et al. (1996, 1998) and Briec and Lesourd (1999), have produced a series of papers in this field. Specifically, Luenberger (1992a, b) introduced the concept of benefit function[1] as a representation of the amount that an individual is willing to trade, in terms of a specific reference commodity bundle. Luenberger also defined a so-called shortage function, which basically measures the distance in the direction of a vector from a production plan (DMU) to the boundary of the production possibility set. In other words, the shortage function measures the amount by which a specific unit is short of reaching the frontier of the production possibility set. Some years later, Chambers et al. (1996, 1998) redefined the benefit function and the shortage function as inefficiency measures, introducing to this end new distance functions, the so called *DDFs*. They showed how the *DDFs* encompass, among others, the Shephard input and output distance functions. And they also derived a dual

[1]Briec and Garderes (2004) have tried to generalize the Luenberger benefit function.

correspondence between the directional distance functions and the profit function that, in their opinion, generalized all previous dual relationships. A few years later, Briec and Lesourd (1999) introduced the so-called Hölder metric distance functions intending to relate the concept of efficiency and the notion of distance in topology. Along these lines, they proved that the profit function can be derived from the Hölder metric distance functions and that these distance functions can be recovered from the profit function.

In contrast to the parametric literature on efficiency, where the measurement of technical efficiency in the context of multiple-outputs is based on a few measures in practice, basically the Shephard input and output distance functions and the directional distance functions, the first years of life of DEA saw the introduction of a bunch of different technical efficiency/inefficiency measures, such as the Russell input and output measures of technical efficiency and their graph extension, the Russell Graph Measure of technical efficiency (see Färe et al. 1985), as well as the additive model (Charnes et al. 1985), followed, several years later by the Range-Adjusted Measure (Cooper et al. 1999), the Enhanced Russell Graph Measure (Pastor et al. 1999) re-baptized as the Slacks-Based Measure (Tone 2001), or the Bounded Adjusted Measure (Cooper et al. 2011a), to name but a few. This short list shows that there is a wide array of tools available for estimating technical inefficiency in the non-parametric world.

On the other hand, most of the classical results and applications in microeconomics related to the measurement and decomposition of overall inefficiency, in terms of technical and allocative inefficiency, are based on the notion of distance function[2] and duality theory. A distance function behaves, in fact, as a technical inefficiency measure when an observation belonging to the corresponding technology is evaluated, with a meaning of 'distance' from the assessed interior point to the boundary of the production possibility set. Also, the distance functions have dual relationships with well-known support functions in microeconomics, as the profit function or the cost and revenue functions, depending on the suppositions that we are willing to assume with respect to the firms´ behavior. In a non-parametric framework, the use of typical parametric tools, such as the Shephard distance functions or the directional distance function, is possible, because their duality relationships with classical support functions were proved for production possibility sets fulfilling general axioms (e.g. convexity) and, in particular, they can be applied to non-parametric polyhedral technologies. Nonetheless, the majority of attempts for estimating overall efficiency have overlooked the concept of distance function, a fact that contrasts significantly with the traditional view of economics of production, where both this concept and duality are the cornerstones of the applied theory. In this respect, some researchers have tried to use additive-type models in DEA for measuring not only technical inefficiency but also profit inefficiency without resorting directly to the notion of distance function (Cooper et al. 2011b and

[2]We would like to remark that moving from the Shephard distance functions to the directional distance functions entails moving from the inverse of efficiency measures to inefficiency measures.

Aparicio et al. 2013). A similar treatment has been applied to Russell oriented-measures (Aparicio et al. 2015), without being able to derive duality results for the corresponding non-oriented measure known as the Enhanced Russell Graph measure (Pastor et al. 1999). However, DEA is a field where there are other alternative efficiency measures and where it seems possible to introduce new ones. Therefore, defining an appropriate methodology to measure and decompose overall inefficiency with whatever DEA measure is something necessary. We accomplish this task in Sects. 2.5–2.7, resorting to the definition of a new concept, the *RDDF*.

We close this chapter by searching, in Sect. 2.8 and within a DEA framework, for an alternative projection for each unit that minimizes profit inefficiency. The new alternative projection has to dominate the point being rated, which guarantees that technical inefficiency can also be evaluated. Finally, we present our conclusions in Sect. 2.9, and add an Appendix for searching for alternative optimal solutions, taking any additive type model as reference.

2.2 Associating a *RDDF* Inefficiency Measure to Any Known DEA Inefficiency Measure

First of all, let us introduce the definition of the traditional *DDF* within a DEA framework. That means that the production possibility set is generated based on a finite sample of units to be rated, and that the inefficiency associated to each unit is obtained by solving a linear program. From now on, we will further assume variable returns to scale (VRS), which guarantees that the three economic functions we are going to consider later on are well-defined.

Let us consider a sample of n units to be rated. Unit $j \in \{1, 2, \ldots, n\}$ uses a specific amount of m inputs, $x_j = \left(x_{1j}, \ldots, x_{mj}\right) \in R_+^m$, to produce a certain amount of s outputs $y_j = \left(y_{1j}, \ldots, y_{sj}\right) \in R_+^s$. As usual, let us denote the unit to be rated as (x_0, y_0).[3] The production possibility set generated by the finite sample of units is

$$T = \left\{ (x, y) \in R_+^{m+s} : \sum_{j=1}^{n} \lambda_j x_{ij} \leq x_i, \forall i, \sum_{j=1}^{n} \lambda_j y_{rj} \geq y_r, \forall r, \lambda_j \geq 0, \forall j, \sum_{j=1}^{n} \lambda_j = 1 \right\},$$

while the efficient frontier of T is defined as[4]

[3]The condition that inputs and outputs need to be non-negative can be relaxed provided the considered *DDF* is translation invariant (see Aparicio et al. 2016).

[4]The efficient frontier, $\partial(T)$, or simply the frontier of T, comprises the weak-efficient frontier. $\partial^W(T)$, and the strong-efficient frontier, or subset of all the Pareto efficient points. See Färe et al. (1985) for the definition of $\partial^W(T)$.

$$\partial(T) := \{(x,y) \in T : \hat{x} \leq x, \hat{y} \geq y \quad and \quad (x,y) \neq (\hat{x},\hat{y}) \Rightarrow (\hat{x},\hat{y}) \notin T\}.$$

Each DDF (Chambers et al. 1996, 1998) is identified by specifying a directional vector $g = (-g_x, g_y) \neq 0_{m+s}$, $g_x \in R^m_+$, $g_y \in R^s_+$. In order to measure the inefficiency associated to a specific unit of the sample, the DDF projects the unit onto the weakly efficient frontier of the technology along the positive semi-ray defined by vector g. Additionally, g may be constant, i.e. g is the same vector for all units, or may be variable, i.e. it is a specific vector for each unit. In the latter case and for unit (x_0, y_0), we write g_0 instead of g. By definition, the projection of unit (x_0, y_0) onto the efficient frontier is the intersection of the semi-ray $\{(x_0, y_0) + \beta_0(-g_{0x}, g_{0y}), \beta_0 \geq 0\}$ with the efficient frontier. The specific value of scalar β_0 that identifies the point of intersection is the inefficiency value measured by the *DDF* associated to point (x_0, y_0), obtained as the optimal solution, β_0^*, of the next linear program, which corresponds to a generic *DDF* working under VRS.[5]

$$\begin{aligned}
\vec{D}(X_0, y_0; g_{0x}, g_{0y}) = Max \quad & \beta_0 \\
s.t & \\
& \sum_{j=1}^{n} \lambda_{j0} X_{ij} \leq X_{i0} - \beta_0 g_{i0x}, \quad i = 1, \ldots, m \\
& \sum_{j=1}^{n} \lambda_{j0} y_{rj} \geq y_{r0} + \beta_0 g_{r0y}, \quad r = 1, \ldots, s \\
& \sum_{j=1}^{n} \lambda_{j0} = 1, \\
& \lambda_{j0} \geq 0, \quad j = 1, \ldots, n
\end{aligned} \tag{2.1}$$

It is well known that $\beta_0^* = 0$ identifies the unit being rated as efficient[6], while $\beta_0^* > 0$ identifies the unit being rated as inefficient.

[5]The linear program we are working with corresponds to the "envelopment form" associated to a *DDF*. Its linear dual is known as the "multiplier form" of the *DDF*. The "envelopment form" deals with units and evaluates their efficient projections, working in the $m + s$ dimensional space where each coordinate corresponds to an input or to an output, while the "multiplier form" identifies the supporting hyperplane of each efficient projection and works also in an $m + s$ dimensional space where each coordinate corresponds to the shadow price of an input or of an output. In this chapter we will only consider "envelopment forms".

[6]An efficient unit is a unit that belongs to the efficient frontier. Any efficient unit may be strongly efficient or, alternatively, weakly efficient. The subset of the efficient frontier of strongly efficient points is called the strongly efficient frontier. The directional vector may also be specified as $(g = g_x, g_y)$. What matters is that g_x appears preceded by a minus sign in (2.1)

Let us now assume that we have obtained the projection[7] of any point of our sample by means of a specific DEA inefficiency model, identified as M. As explained in Footnote 7, if M is single-valued, the projection of any point is unique. Although the method we are going to propose is valid for any DEA model, the most usual case is that inefficiency model M is a linear programming model or a programming model that can be linearized through an appropriate change of variables[8] (see, e.g., Pastor et al. 1999).

If (x_0, y_0) denotes the point being rated, let us denote as $\left(x_0^M, y_0^M\right)$ its efficient projection. Moreover, let us denote as Π^M the set of efficient projections obtained through M and associated to the sample of points being rated.[9] Let us further denote as $\tau_0^M := TI^M(x_0, y_0)$ the technical inefficiency evaluated by means of model M. It is well known that $\tau_0^M \geq 0$. Moreover, if (x_0, y_0) belongs to the efficient frontier, then $\tau_0^M = 0$.

Definition 1 Associated to DEA model M and to the evaluated set of efficient projections, Π^M, we define the *reverse directional distance function model*,$RDDF^{M,\Pi^M}$, by specifying the directional vector g_0^{M,Π^M} at point (x_0, y_0) as follows[10]:

$$g_0^{M,\Pi^M} := \left\{ \begin{array}{ll} \frac{1}{\tau_0^M}\left(x_0 - x_0^M, y_0^M - y_0\right) \geq 0_{m+s}, & \text{if} \quad \left(x_0 - x_0^M, y_0^M - y_0\right) \neq 0 \text{ and } \tau_0^M > 0 \\ (1_m, 1_s), & \text{if} \quad \tau_0^M = 0 \end{array} \right\}. \tag{2.2}$$

As usual, the technical inefficiency associated to the directional distance function $RDDF^{M,\Pi^M}$ is denoted as $\vec{D}\left(x_0, y_0; g_{0x}^{M,\Pi^M}, g_{0y}^{M,\Pi^M}\right)$. The definition of g_0^{M,Π^M} distinguishes between two kinds of points, just as model M does: the points that get an

[7]Usually DEA researchers and practitioners are satisfied computing a unique projection for each point, although in many DEA models multiple projections can be identified. Here we accommodate our findings to this tradition and work initially with a single projection for each point. Nevertheless, in Sect. 2.4, Example 4.2, we consider an input-oriented additive model with multiple projections. DEA models with a unique projection for each point can be baptized as "single-value DEA models", as opposed to "multiple-value DEA models".

[8]If M is not a DEA inefficiency model but a DEA efficiency model, we can always conveniently modify its objective function so as to get an inefficiency model (see, e.g., Aparicio et al. 2015). The novel loss distance function (Pastor et al. 2012) embraces all well-known DEA models as they are or with minor changes in its objective function, and constitutes the widest known family of DEA inefficiency models. Since the loss distance function considers the multiplier form of each DEA model we will not go into further details.

[9]Observe that we obtain a single projection for each point through the corresponding linear program. This fact does not exclude the possible existence of alternative optimal projections, whose study is introduced only for additive type models in Appendix 1.

[10]The name "reverse" directional distance function (*RDDF*) is a consequence of how we define the associated *DDF*. Usually, for defining a *DDF*, we need to know, at each point, the corresponding directional vector and, based on it, we determine its efficient projection. In our new proposed approach, we do it the other way round, i.e., we know beforehand the projection of each point being rated and, based on it, we derive the corresponding directional vector at that point.

inefficiency score $\tau_0^M > 0$ and the rest of the points, for which $\tau_0^M = 0$. In any case, the second subset of points that satisfy $\tau_0^M = 0$ corresponds to all the points whose projection through model M belongs to the frontier. It is clear that for any point of this second subset, the proposed fix directional vector $(1_m, 1_s)$ of $RDDF^{M,\Pi^M}$ will assign an inefficiency equal to 0, just as model M does.

As a direct consequence of the last definition the next statements holds.

Proposition 1

(a) *$RDDF^{M,\Pi^M}$ has exactly the same projections as M.*[11] *As a consequence, $RDDF^{M,\Pi^M}$ inherits the same frontier and the same returns to scale characteristics as M.*

(b) *The* technical inefficiency *associated to any unit (x_0, y_0) through $RDDF^{M,\Pi^M}$ is exactly the same as the technical inefficiency evaluated through M, τ_0^M.*

Proof

(a) Trivial.

(b) We have two cases. First, if $\tau_0^M = 0$ then (x_0, y_0) belongs to the efficient frontier and, by (2.2), $g_0^{M,\Pi^M} = (1_m, 1_s) > 0_{m+s}$. Therefore, $\vec{D}\left(x_0, y_0; g_{0x}^{M,\Pi^M}, g_{0y}^{M,\Pi^M}\right) = 0$. Second, if $\tau_0^M > 0$, by Lemma 2.2(c) in Chambers et al. (1998) we have that $\vec{D}\left(x_0, y_0; g_{0x}^{M,\Pi^M}, g_{0y}^{M,\Pi^M}\right) = \tau_0^M \vec{D}\left(x_0, y_0; \left(x_0 - x_0^M\right), \left(y_0^M - y_0\right)\right)$ $= \tau_0^M$, as a consequence of being $\vec{D}\left(x_0, y_0; \left(x_0 - x_0^M\right), \left(y_0^M - y_0\right)\right) = 1$. ∎

As a direct consequence of Proposition 1(a), we get the next Corollary.

Corollary 1.1 *If model M projects all units onto the strong (weak) efficient frontier, so does $RDDF^{M,\Pi^M}$.*

Corollary 1.1 shows an easy way to generate DDFs with all their projections onto the strongly efficient frontier,[12] something that seldom occurs in the framework of the usual directional distance functions.

The next result is straightforward and adds consistency to our proposal.

[11] We would like to remind the reader that the definition of the *RDDF* is based on the evaluation of a single projection for each of the points being rated. In case one or more points have multiple possible projections, the change of the projection of any of the considered points gives rise to a different associated *RDDF*. This is the reason for writing $RDDF^{M,\Pi^M}$, making explicit through the super-indexes that *RDDF* depends not only on model M but on the set of computed projections Π^M.

[12] This happens exactly when M delivers a strong inefficiency measure, a measure whose projections belong to the strongly efficient frontier. A well-known example is the weighted additive model (Lovell and Pastor 1995) that will be considered in the examples of the next three sections.

Corollary 1.2 *If model M corresponds to a DDF, then M and* $RDDF^M$ *collapse together, with the possible exception of the directional vectors associated to points being rated where* $\tau_0^M = 0$.

Proof According to expression (2.2) it is obvious that at any point being rated where $\tau_0^M > 0$, both *M* and $RDDF^M$ are exactly the same *DDF*. Moreover, it is also obvious that at any point being rated where $\tau_0^M = 0$, the *DDF* directional vector associated to that point may be different from $(1_m, 1_s)$, which is the fixed directional vector that expression (2.2) assigns to that point in the definition of $RDDF^M$. ■

The next section takes advantage of the above introduced $RDDF^{M,\Pi^M}$ allowing us to transform any "weak *DDF*", i.e., any *DDF* whose set of efficient projections does not belong to the strongly efficient frontier, into a closely related "strong *DDF*", that is, a *DDF* whose set of efficient projections belongs to the strongly efficient frontier. In many real life applications, non-dominated peers are preferred over dominated ones, which is precisely the difference between strongly efficient projections and non-strongly efficient ones. This is the main reason for introducing Sect. 2.3.[13] This was also the objective of Fukuyama and Weber (2009), where a modified directional distance function was defined. As Pastor and Aparicio (2010) pointed out, this modified *DDF* really coincides with a specific weighted additive measure.

2.3 Generating a Strong DDF Based on a Weak DDF

We know in advance that all the projections associated to any *DDF* definitely belong to the efficient frontier and, sometimes, they all belong to the strongly efficient frontier. Let us start showing this last unusual case by means of an example.

Example 3.1 Analyzing a Strong *DDF*

Let us consider, in the one input—one output space, the next set of units to be rated U1(2,4), U2(4,8), U3(6,2), U4(3,4) and U5(10,8). The corresponding directional vectors are: (4,1) for U1, (8,2) for U2, (4,2) for U3, (4,1) for U4 and (1,0) for U5. As usual, we assume a VRS technology. After performing the first stage we get the corresponding set of projections (see Table 2.1). The results of Table 2.1 suggest that U1 and U2 are strongly efficient units.[14] U3 is an inefficient unit that belongs to

[13]Although we focus here on *DDF* inefficiency measures, it is easy to associate a strong *DDF* to any DEA inefficiency measure by resorting, as we do here, to the *RDDF*, as explained in more detail in Sect. 2.4.

[14]This fact can be corroborated by means of a two-dimensional graphical display.

Table 2.1 Results associated with Example 3.1

Unit: (x_0, y_0)	Directional vector	Inefficiency β_0^*	Projection: $\left(x_0^{p1}, y_0^{p1}\right)$	Input slack: s^{-*}	Output slack: s^{+*}
U1(2,4)	(4,1)	0	(2,4) = U1	0	0
U2(4,8)	(8,2)	0	(4,8) = U2	0	0
U3(6,2)	(4,2)	1	(2,4) = U3 + 1(−4,2) = U1	0	0
U4(3,4)	(4,1)	2/9	(19/9,38/9) = U4 + 2/9 (−4,1) = 17/18U1 + 1/18U2	0	0
U5(10,8)	(1,0)	6	(4,8) = U5 + 6(−1,0) = U2	0	0

the interior of the production possibility set and whose projection (2,4) is the strongly efficient point U1.

U4 is also an interior point whose *DDF* projection is a strongly efficient point that is a linear convex combination of the two strongly efficient units:

$$U4 + \frac{2}{9}(-4,1) = \left(3 - \frac{8}{9}, 4 + \frac{2}{9}\right) = \left(\frac{19}{9}, \frac{38}{9}\right) = \frac{17}{18}U1 + \frac{1}{18}U2.$$

Finally, U5 is itself a weakly efficient point dominated by its projection U2:

$$U5 + 6(-1,0) = (10,8) - (6,0) = (4,8) = U2.$$

In this simple example, we have been able to identify each of the two strongly efficient units, U1 or U2. In this particular case a convex combination of U1 and U2 is also a strongly efficient point, as for example the projection of U4, which means that all the located projections belong to the strongly efficient frontier.

Let us now consider the most frequent case, i.e., a weak *DDF*, where at least one of the projections of the sample of points being rated does not belong to the strongly efficient frontier. Let us first introduce a procedure for classifying the projection of any of the weak *DDF* inefficient points as strongly efficient or as not strongly efficient. In the second case, the procedure we are going to introduce is able to identify a new strongly efficient point that dominates the initial weakly efficient projection.

2.3.1 Converting a Weak DDF into a Strong DDF

Some years ago Asmild and Pastor (2010) designed a two stage procedure for getting strongly efficient projections in two particular cases: the multi-directional analysis measure (MEA) of Bogetoft and Hougaard (1999), and the range directional measure (RDM) of Silva-Portela et al. (2004). Although both are efficiency measures, the same reasoning can be applied to inefficiency measures, such as *DDFs*. We are going to replicate the procedure here for analyzing any *DDF*

(Chambers et al. 1998). Basically the first stage will get the projections of the considered *DDF* model, and the second stage will project the obtained projection onto the strongly efficient frontier, with the help of the additive model (Banker et al. 1984), formulated as follows:

$$
\begin{aligned}
Add(x_0, y_0) = Max_{s^-, s^+, \lambda} \quad & \sum_{i=1}^{m} s_{i0}^- + \sum_{r=1}^{s} s_{r0}^+ \\
s.t. \quad & \\
& \sum_{j=1}^{n} \lambda_j x_{ij} = x_{i0} - s_{i0}^-, \quad i = 1, \ldots, m \\
& \sum_{j=1}^{n} \lambda_j y_{rj} = y_{r0} + s_{r0}^+, \quad r = 1, \ldots, s \\
& \sum_{j=1}^{n} \lambda_j = 1, \\
& \lambda_j \geq 0, \quad j = 1, \ldots, n \\
& s_{i0}^- \geq 0, \quad i = 1, \ldots, m, \quad s_{r0}^+ \geq 0, \quad r = 1, \ldots, s
\end{aligned}
\tag{2.3}
$$

Additive model (2.3) has the next advantage over other used DEA models, such as radial models or DDF models: it always achieves a strongly efficient projection for any point being rated. Since the additive model identifies an L_1-path towards the frontier connecting the point being rated and its projection, and the length of this path is obtained as the sum of all the optimal slack values that appear in the objective function, named as $Add(x_0, y_0)$ in model (2.3), it is straightforward to enunciate the next statement: "(x_0, y_0), the point being rated, is a strongly efficient point if, and only if, $Add(x_0, y_0) = 0$".[15]

Now, let us go back to *DDF* model (2.1). We can reformulate it very easily by taking the following action: add a single slack variable to each inequality transforming it into an equality as follows.

[15]The subset of strongly efficient point of the sample being rated are denoted as *E*. There are many other strongly efficient points of the production possibility set that do not belong to *E*. Under VRS, only convex linear combinations of points of *E* are potential candidates. Again, in order to check if one of these points is strongly efficient or not, the easiest way is to resort to the additive model [3] and analyze the mentioned convex linear combination of points of *E*. Only if the optimal objective value is 0, or, equivalently, all the optimal slack values are 0, the point being rated is strongly efficient.

$$\begin{aligned}\vec{D}(x_0, y_0; g_{0x}, g_{0y}) = Max \quad & \beta_0 \\ s.t \quad & \\ & \sum_{j=1}^{n} \lambda_j x_{ij} + s_{i0}^{-} = x_{i0} - \beta_0 g_{i0x}, \quad i = 1, \ldots, m \\ & \sum_{j=1}^{n} \lambda_j y_{rj} - s_{r0}^{+} = y_{r0} + \beta_0 g_{r0y}, \quad r = 1, \ldots, s \\ & \sum_{j=1}^{n} \lambda_j = 1, \\ & \lambda_j \geq 0, \qquad\qquad\qquad\qquad j \in E\end{aligned} \tag{2.4}$$

We are searching for strongly efficient projections. Therefore, we assume that the projection identified through (2.4) is no longer $(x_{01} - \beta_0^* g_{01x}, \ldots, x_{0m} - \beta_0^* g_{0mx}, y_{01} + \beta_0^* g_{01y}, \ldots, y_{0s} + \beta_0^* g_{0sy})$ as in model (2.1), but $\sum_{j=1}^{n} \lambda_{j0}^* (x_j, y_j)$, a point that satisfies

$$\sum_{j=1}^{n} \lambda_j^* (x_j, y_j) + (s_0^{-*}, -s_0^{+*}) = (x_0, y_0) + \beta_0^* (-g_{0x}, g_{0y}). \tag{2.5}$$

This constitutes the basic difference between model (2.4) and (2.1). As a direct consequence of (2.5), it is clear that $\sum_{j=1}^{n} \lambda_{j0}^* (x_j, y_j)$ dominates, or is equal, to $(x_0, y_0) + \beta_0^* (-g_{0x}, g_{0y})$, and, consequently, is closer to the strongly efficient frontier. Our proposed second stage, based, as said before, on the additive model, checks if the first stage projection $\sum_{j=1}^{n} \lambda_{j0}^* (x_j, y_j)$ is itself a strongly efficient point or, alternatively, finds a strongly efficient point for replacing it.

As explained before, we are going to design a *two stage process,* which combines a given *DDF*, which offers us a first stage projection for each point of the sample, with a second stage *additive model*, that projects each first stage projection onto the strongly efficient frontier, ending up with a second stage strongly efficient projection. Relating each inefficient point with its final second stage projection gives rise to a comprehensive *DDF* inefficiency measure that combines all the detected inefficiencies into a single number.

Second Stage Analysis: Identifying the Strongly Efficient Projections

In order to classify the first stage projection $\left(x_0^{p1}, y_0^{p1}\right)$ of (x_0, y_0) obtained through model (2.4), and as a direct consequence of (2.5) we can write the next equivalent expression:

$$\sum_{j=1}^{n} \lambda_j^* \left(x_j, y_j\right) = (x_0, y_0) + \beta_0^* (-g_{0x}, g_{0y}) + \left(-s_0^{-*}, s_0^{+*}\right). \tag{2.6}$$

Take additive model (2.3) and evaluate point $\sum_{j=1}^{n} \lambda_j^* \left(x_j, y_j\right)$. The result can be written as

$$\sum_{j=1}^{n} \lambda_j^* \left(x_j, y_j\right) = \sum_{j=1}^{n} \hat{\lambda}_j \left(x_j, y_j\right) + \left(\hat{s}_0^-, -\hat{s}_0^+\right), \tag{2.7}$$

where $\left(\hat{s}_0^-, \hat{s}_0^+\right)$ is the optimal solution of (2.3) linked to $\hat{\lambda}$. We know that $\sum_{j=1}^{n} \hat{\lambda}_j \left(x_j, y_j\right)$ is a strongly efficient point identified through our additive second stage projection and named alternatively as $\left(x_0^{p2}, y_0^{p2}\right)$. Combining (2.6) with (2.7) we can relate our initial inefficient point (x_0, y_0) with our second stage projection, obtaining the next relationship:[16]

$$\begin{aligned} \left(x_0^{p2}, y_0^{p2}\right) &= \sum_{j=1}^{n} \hat{\lambda}_j \left(x_j, y_j\right) = \sum_{j=1}^{n} \lambda_j^* \left(x_j, y_j\right) - \left(\hat{s}_0^-, -\hat{s}_0^+\right) \\ &= (x_0, y_0) + \beta_0^* (-g_{0x}, g_{0y}) + \left(-s_0^-, s_0^+\right) - \left(\hat{s}_0^-, -\hat{s}_0^+\right) \\ &= (x_0, y_0) + \beta_0^* (-g_{0x}, g_{0y}) + \left(-\left(s_0^- + \hat{s}_0^-\right), s_0^+ + \hat{s}_0^+\right). \end{aligned} \tag{2.8}$$

In summary, the calculated second stage strongly efficient projection is reached after performing a directional vector movement, $\beta_0^* (-g_{0x}, g_{0y})$, followed by a non-directional L_1-movement,[17] $\left(-\left(s_0^- + \hat{s}_0^-\right), s_0^+ + \hat{s}_0^+\right)$. The compound movement towards the strongly efficient frontier is given by $\beta_0^* (-g_{0x}, g_{0y}) + \left(-\left(s_0^- + \hat{s}_0^-\right), s_0^+ + \hat{s}_0^+\right)$, where the minus and plus signs indicates that in order to reach the strongly efficient frontier we need to reduce inputs and to increase outputs. Let us show an example before assigning a sensible inefficient value to each new strongly efficient projection.

Example 3.2 Part 1. Getting Strong Projections for a Weak *DDF*

Let us consider in the two-input one-output space the next five units: U1(4,2,4), U2(4,4,7), U3 (4,6,9), U4(4,3,2), U5(10,3,1) and U6(8,5,20/3). Let us assume that the considered *DDF* has a unique constant directional vector $g = (2,0,0)$. The first

[16] Even if $\left(s_0^-, s_0^+\right) = 0_{m+s}$, it is not assured that our first stage projection, $\sum_{j=1}^{n} \lambda_j^* \left(x_j, y_j\right)$, is itself a strongly efficient point. What is additionally needed is that the second-stage slacks, $\left(\hat{s}_0^-, \hat{s}_0^+\right)$, are equal to zero. Hence, the second stage is always needed.

[17] This is the typical movement associated with the additive model; the L_1-distance is also known as the Manhattan metric. Moreover, although the considered second stage could have alternative optimal solutions, only one of them is being considering here. Hence we can simplify the notation for describing the associated *RDDF*.

Table 2.2 Results associated with Example 3.2, Part 1

Unit: (x_0, y_0)	Dir. vector	Inefficiency β_0^*	Projection: $\left(x_0^{p1}, y_0^{p1}\right)$	Input 1 slack: $\hat{s}_1^{-*}$	Input 2 slack: $\hat{s}_2^{-*}$	Output slack: $\hat{s}^{+*}$
U1(4,2,4)	(2,0,0)	0	U1	0	0	0
U2(4,4,7)	(2,0,0)	0	U2	0	0	0
U3(4,6,9)	(2,0,0)	0	U3	0	0	0
U4(4,3,2)	(2,0,0)	0	U4	0	0	0
U5(10,3,1)	(2,0,0)	3	U4	0	0	1
U6(8,5,20/3)	(2,0,0)	2	$\frac{14}{15}U2 + \frac{1}{15}U4 = \left(4, \frac{59}{15}, \frac{20}{3}\right)$	0	$\frac{16}{15}$	0

stage projections are reported in Table 2.2. (All our linear programs have been solved resorting to Excel-Solver.) Our *DDF* model projects the first four units onto themselves, which means that, in each case, the directional inefficiency is $\beta^* = 0$ and the three optimal slack values detected by model (2.4) are equal to 0. The projection of U5 is point U4, getting an optimal directional inefficiency equal to 3 ($\beta_{U5}^* = 3$), and optimal slack values equal to (0,0,1). Finally, the optimal value of *DDF* model (2.4) for U6 equals $\beta_{U6}^* = 2$ while the three optimal slack values are $\left(0, \frac{16}{15}, 0\right)$ and its projection is point $\frac{14}{15}U2 + \frac{1}{15}U4$. Hence the first stage projection of U5 is point U4 (4,3,2), and that corresponding to U6 is point (4,59/15,20/3). We further need to determine how the additive model (2.3) rates the six first stage projections (see Table 2.3).

Table 2.3 Results associated with Example 3.2, Part 2

Unit: $\left(x_0^{p1}, y_0^{p1}\right)$	Inefficiency $s_1^{-*} + s_2^{-*} + s^{+*}$	Projection: $\left(x_0^{p2}, y_0^{p2}\right)$	Input1 slack: $\hat{s}_1^{-*}$	Input2 slack: $\hat{s}_2^{-*}$	Output slack: $\hat{s}^{+*}$
U1(4,2,4)	0	U1	0	0	0
U2(4,4,7)	0	U2	0	0	0
U3(4,6,9)	0	U3	0	0	0
U4(4,3,2)	0	0.5(U1 + U2) = $\left(4,3,\frac{11}{2}\right)$	0	0	3.5
U4(4,3,2)	3	0.5(U1 + U2)	0	0	3.5
$\frac{14}{15}U2 + \frac{1}{15}U4 = \left(4, \frac{59}{15}, \frac{20}{3}\right)$	2	$\frac{1}{30}U1 + \frac{29}{30}U2 = \left(4, \frac{59}{15}, \frac{69}{10}\right)$	0	0	$\frac{7}{30}$

The result is that U1, U2 and U3 are rated as strongly efficient points, while the second stage projection of U4 is point 0.5(U1 + U2) = (4,3,11/2), with second stage optimal slacks equal to (0,0,7/2). Hence the final projection of U5 is also point

0.5(U1 + U2), with total inefficiencies equal to 3(2,0,0) +(0,0,7/2) = (6,0,7/2). Finally, the first stage projection of U6 has 1/30(U1 + 29U2) = (4,59/15,69/10) as second stage projection, with second stage optimal slacks equal to (0,0,7/30). Hence the total slacks associated to U6 are 2(2,0,0) +(0,16/15,0) +(0,0,7/30) = (4,32/30,7/15). These results are reported in Table 2.3. In this last case, the presence of second stage slacks indicates that the first stage projection of U6 is not a strongly efficient point.

2.3.2 *Measuring the Comprehensive Inefficiencies of the Derived Strong* DDF

So far we have devised a procedure that generates, after a second stage analysis, a strong *DDF*, whose final outcomes are strongly efficient projections for each unit being rated. This constitutes a simple way of transforming a weak *DDF*, that is, an inefficiency measure that accounts only for directional inefficiencies, into a strong *DDF*, that is, a comprehensive inefficiency measure, that accounts for all types of inefficiencies, both directional and non-directional. In order to complete our proposal, we need to measure the global inefficiency associated to each point. Since, at each point, the inefficiency associated to the strongly efficient projection we are seeking should be comparable to and greater than the inefficiency associated to the initial weakly efficient projection, β_0^*, we propose to normalize the "strong" directional vector, $\left(x_0 - x_0^{p2}, y_0^{p2} - y_0\right)$, obtained through the two stage procedure, with respect to the directional vector considered at the first stage, $\left(x_0 - x_0^{p1}, y_0^{p1} - y_0\right)$. The length of the strong directional vector $\left(x_0 - x_0^{p2}, y_0^{p2} - y_0\right)$ is always greater or equal than the length of the initial directional vector $\left(x_0 - x_0^{p1}, y_0^{p1} - y_0\right)$, because it holds that $\left(x_0 - x_0^{p2}, y_0^{p2} - y_0\right) = \left(x_0 - x_0^{p1}, y_0^{p1} - y_0\right) + \left(x_0^{p1} - x_0^{p2}, y_0^{p2} - y_0^{p1}\right)$ with $x_0^{p2} \leq x_0^{p1}$ and $y_0^{p2} \geq y_0^{p1}$. If $\|g_0\|_2$ denotes the Euclidean norm of Stage I directional vector g_0 and, as said before, we want to normalize the strong directional vector $\left(x_0 - x_0^{p2}, y_0^{p2} - y_0\right)$ so as to get the same "length", all we have to do is to divide it by its actual length $\left\|\left(x_0 - x_0^{p2}, y_0^{p2} - y_0\right)\right\|_2$ and to multiply it by $\|g_0\|_2$. As a consequence, the associated strong inefficiency value β_0^{*p2} satisfies

$$\beta_0^{*p2} := \frac{\left\|\left(x_0 - x_0^{p2}, y_0^{p2} - y_0\right)\right\|_2}{\|g_0\|_2}, \tag{2.9}$$

which, as explained above, is at least as big as $\beta_0^* = \dfrac{\left\| \left(x_0 - x_0^{p1}, y_0^{p1} - y_0 \right) \right\|_2}{\|g_0\|_2}$.[18]

Example 3.2 Part 2. Estimating the Inefficiencies of the Derived Strong *DDF*.

Let us consider again the data and the results of Example 3.2. In short, at the first stage, the considered *DDF* rated the first four units as directional efficient. At the second stage, only the first three units remained truly efficient—U1(4,2,4), U2(4,4,7), U3 (4,6,9)—, while the fourth one, U4(4,3,2), the fifth one U5(10,3,1) and the sixth one U6(8,5,20/3) were rated as inefficient and achieved second stage projections different from the first stage. Moreover, the strongly efficient projection of U4 and U5, after performing the second stage, was point (4,3,11/2). Finally, the second stage strongly efficient projection of U6 was point (4, 59/15, 69/10). Hence, the three directional vectors that connect U4, U5 and U6 with their final projections that were calculated at the bottom of Example 3.2, were $\left(0,0,\frac{7}{2}\right)$, $\left(-6,0,\frac{7}{2}\right)$, and $\left(-4,-\frac{16}{15},\frac{7}{30}\right)$. It is easy to verify that their corresponding Euclidean norms are 3.5, 6.946 and 4.146, respectively. In order to standardize the three new evaluated directional vectors with respect to the unique directional vector of Stage I, vector (2,0,0), we need to shorten them to reduce their length to 2, which means dividing each of them by its length and, at the same time, multiplying each of them by 2. Since the original connecting vector detects always an inefficiency equal to 1, the new shortened reverse directional vector will exhibit an inefficiency value of $\frac{3.5}{2} = 1.75$ for U4, $\frac{6.946}{2} = 3.473$ for U5, and $\frac{4.146}{2} = 2.073$, for U6. These new comprehensive inefficiencies are higher than the directional inefficiency reported in Table 2.2 (0, 3, and 2, respectively), and correspond to the inefficiency value expression contained in (2.9). Hence, the new strong *DDF* detects an inefficiency higher than the weak *DDF* for the three considered inefficiency units, which is reasonable because the new *DDF* accounts for all types of inefficiencies, the directional inefficiency as detected by the weak *DDF* and the inefficiencies detected by the slacks, which are basically non-directional and are quite often greater than 0. The interesting point is that all these inefficiencies can be combined and measured through a new strong *DDF* that, obviously, avoids slacks, and offers a single number that measures all types of inefficiencies.

2.4 Deriving DEA Comprehensive Inefficiency Measures. the Case of the Comprehensive Radial Models

Thanks to the exercise performed in Sect. 2.3 it is easy to build a comprehensive inefficiency measure based on any weak DEA inefficiency model. In fact, any model *M* gives rise to the corresponding $RDDF^{M,\Pi^M}$. Considering this *RDDF* as the

[18]Observe that each component of the vector in the numerator equals the corresponding component of the vector in the denominator times β_0^*.

DDF of the first stage and applying to it the findings of Sect. 2.3, namely the two stage procedure, we end up creating a new strong *RDDF*, which is exactly the solution we are looking for. Based on the results of Sect. 2.3 it is straightforward to formulate the directional vector that relates model *M* with its strong *RDDF*, a new expression that substitutes the expression given by (2.2). Let us name the new directional vector as $\hat{g}_0^M$. Now the distinction is not based on the original directional inefficiency values $\tau_0^M \geq 0$, but on the comprehensive inefficiency values obtained after performing the second stage, denoted as T_0^M. It is obvious that $T_0^M \geq \tau_0^M$. Let us observe that only the strongly efficient points will get $T_0^M = 0$, which means that any non-strongly efficient point will get $T_0^M > 0$. Hence, new expression (2.2) will be based on *E*, the subset of strongly efficient points, and not on τ_0^M. The new expression that gives the directional vector associated to each point (x_0, y_0) being rated and that relates model *M* with the corresponding derived strong *RDDF* is:

$$\hat{g}_0^M := \left\{ \begin{array}{ll} \frac{1}{T_0^M}\left(x_0 - x_0^{p2}, y_0^{p2} - y_0\right) \geq 0_{m+s}, & if\ (x_0, y_0) \notin E \\ (1_m, 1_s), & if\ (x_0, y_0) \in E \end{array} \right\} \tag{2.10}$$

Let us observe that the strong *RDDF* represents the comprehensive inefficiency model associated to model *M*. Hence, our procedure is a general procedure for generating comprehensive inefficiency models based on any DEA inefficiency model.[19]

In particular, we are ready to define comprehensive Radial Models. These particular models are even easier to deal with, because we know how to formulate them as DDFs and, therefore, Sect. 2.3 gives us the solution directly. The constant returns to scale (CRS) Radial Models were the first defined DEA models and, in honor of their authors, they are known as the CCR models (Charnes et al. 1978). Later on, the VRS Radial Models were introduced (Banker et al. 1984) and, for the same mentioned reason, they are also known by the acronym BCC. We will focus our attention on the latter ones. There are two BCC models: the BCC input-oriented and the BCC output-oriented models. The first one is formulated as follows.[20]

[19]We would like to point out that a referee has pointed out that our procedure opens a new research avenue for ranking units resorting to appropriate *DDFs*, by selecting appropriate directional vectors.

[20]As already mentioned in Footnote 5, we will only consider "the envelopment form" of any BCC model and will not even mention its linear dual program known as "the multiplier form", which, in this case, can also be expressed as a linear fractional program called "the ratio form".

$$
\begin{aligned}
BCC^{IO}_{EFF}(x_0, y_0) = Min \quad & \theta_0 \\
& s.t. \\
& \sum_{j \in E} \lambda_j x_{ij} \leq \theta_0 x_{i0}, \quad i = 1, \ldots, m \\
& \sum_{j \in E} \lambda_j y_{rj} \geq y_{r0}, \quad r = 1, \ldots, s \\
& \sum_{j \in E} \lambda_j = 1, \\
& \lambda_j \geq 0, \quad j = 1, \ldots, n
\end{aligned}
\tag{2.11}
$$

This model is a DEA efficiency model, and θ_0^* is known as the input efficiency score of unit (x_0, y_0). Being aware that θ_0^* is a quantity in between 0 and 1, making the change of variable $\theta_0 = 1 - \beta_0$ we just move from an efficiency model to an inefficiency model. Since minimizing θ_0 is equivalent to maximizing $-\theta_0 = \beta_0 - 1$, the last model can be reformulated as the next DDF model,[21] provided we add slack variables in the same way as we did in (2.4) for model (2.1).

$$
\begin{aligned}
BCC^{IO}_{INEFF}(x_0, y_0; x_0, 0_s) \quad & = Max \quad \beta_0 \\
& s.t. \quad \sum_{j \in E} \lambda_j x_{ij} + s^-_{i0} = x_{i0} - \beta_0 x_{i0}, \quad i = 1, \ldots, m \\
& \sum_{j \in E} \lambda_j y_{rj} - s^+_{r0} = y_{r0}, \quad r = 1, \ldots, s \\
& \sum_{j \in E} \lambda_j = 1, \\
& \lambda_j \geq 0, \qquad j = 1, \ldots, n \\
& s^-_{i0}, s^+_{r0} \geq 0, \quad i = 1, \ldots, m, \\
& \qquad\qquad\quad r = 1, \ldots, s
\end{aligned}
\tag{2.12}
$$

It is obvious that $\beta_0^* = 1 - \theta_0^* \geq 0$ because $0 \leq \theta_0^* \leq 1$. Moreover, $\beta_0^* \leq 1$. Since the directional vector $(x_0, 0_s)$ is unable to change outputs, it is input-oriented. Moreover, since the change in inputs is guided by x_0 we pursue a proportional—or radial—reduction in inputs. It is well established that this model does not necessarily project all the inefficient points onto the strongly efficient frontier, or, in other words, depending on the sample of units being rated, BCC^{IO}_{INEFF} can be a weak *DDF*. Applying to it the two stage procedure developed in Sect. 2.3 it is easy to end

[21] Strictly speaking, model $BCC^{IO}_{INEFF}(x_0, y_0; x_0, 0_s)$ has as objective function $\beta_0 - 1$ instead of β_0, but its expression as a *DDF* requires the proposed related objective function.

up with a comprehensive strong *DDF*, as well as with the corresponding derived strong inefficiency scores. Although the initial weak inefficiency scores are always less than or equal to 1, the final strong inefficiency scores may be greater than 1.

Example 4.1 Deriving a Comprehensive BCC Input-Oriented Model

Let us consider in the two-input one-output space the next five units: U1(4,2,40), U2(4,4,90), U3(4,6,120), U4(40,30,4) and U5(4,4,80). Applying model (2.12) we ensure that the BCC input-oriented inefficiency model rates the first three units as efficient, assigning each of them an optimal inefficiency of $\beta^* = 0$ and all optimal slack values at level 0. Furthermore, model (2.12) assigns to U4 an optimal inefficiency value $\beta^*_{U4} = 0.9$, and optimal slack values equal to (0,1,36). Consequently, the first stage projection of U4 is point (40,30,4) − 0.9(40,30,0) + (0, −1,36) = (4,2,40), which is exactly U1. Finally, U5 is projected onto 0.5 (U1 + U3) = U5, which means that $\beta^* = 0$ and all the optimal slacks are at level 0. We need to perform Stage 2 since we are not sure which of the projections are strongly efficient. First we check, by means of the additive model, if U1, U2, and U3 are strongly efficient units. The answer is yes, and, as a consequence, the first stage projections of U1, U2, U3, and U4 are also its second stage projections. Hence, U4 will maintain its directional inefficiency value as well as its slack non-directional inefficiency value. Both types of inefficiencies will be jointly considered when evaluating the corresponding strong reverse directional vector. Finally, U5 gets a new second stage projection, precisely U2, with the only non-zero slack on the output side and equal to 10. Again, the strong reverse directional vector associated to the second stage projections are, according to (2.10), vector (1,1;1) for U1, U2 and U3, with associated inefficiency equal to 0. The reverse directional vector for U4 is obtained as U1 − U4 = (4,2,40) − (40,30,4) = (−36, −28, 36), while the associated to U5 is U2 − U5 = (4,4,90) − (4,4,80) = (0,0,10). In the last two cases the associated comprehensive inefficiency is obtained as explained before. The Euclidian length of (−36, −28, 36) is 58.103 and the corresponding to (0,0,10) is 10. Since, associated to the first stage projection the used directional vectors were (40,30,0) for U4 and (4,4,0) for U5, their lengths were 50 and 5.657. Therefore, the comprehensive inefficiency of the strong *RDDF* are $\frac{58.103}{50} = 1.162 > 1$ for U4 and $\frac{10}{5.697} = 1.755 > 1$ for U5.

This example shows that non-radial inefficiencies may be relevant enough so as to get an inefficiency score bigger than 1, and that the original input orientation has been lost because the final strong directional vector at U4 has all positive components, which means that it modifies all inputs and outputs in order to reach the desired strong projection. Curiously enough, for U5, its final strong directional vector is output oriented.

Even in the simplest one input-one output space the second stage may be necessary, as shown in the next example.

Example 4.2 The BCC Input-Oriented Model with just Two Dimensions.

In the one input—one output space let us consider the next sample of 6 units: (2,2), (4,2), (2,10), (4,14), (5,14) and (12,12). If we draw a picture, see Fig. 2.1, and apply the BCC input-oriented inefficiency model, as given by (2.12), we get the results listed in Table 2.4.

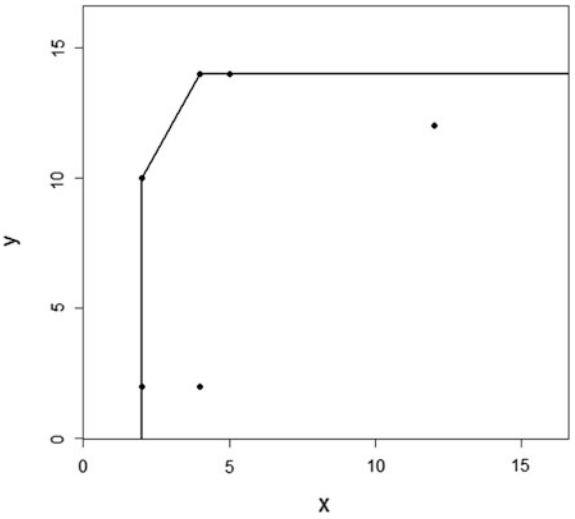


Fig. 2.1 Figure associated with Example 4.2

Table 2.4 Results associated with Example 4.2, First Stage

Unit	β_0^*	$\sum_{j \in E} \lambda_j^* (x_j, y_j)$	Input slack	Output slack	Unit being rated is input-radial
U1(2,2)	0	U1	0	0	Efficient
U2(4,2)	1	U1	0	0	Inefficient
U3(2,10)	0	U3	0	0	Efficient
U4(4,14)	0	U4	0	0	Efficient
U5(8,14)	2	U4	0	0	Inefficient
U6(12,12)	4.5	0.5(U3 + U4) = (3,12)	0	0	Inefficient

Let us observe that the first stage does classify each unit as input-radial efficient or inefficient—see last column of Table 2.4—but is unable to tell us if the radial efficient units belong to the strongly efficient frontier or not. Moreover, it is also unable to tell us if the input-radial inefficient units belong to the interior of the production possibility set or to the weak part of the efficient frontier. For getting this information we need to perform our second stage analysis, whose results are listed in Table 2.5.

Table 2.5 Combined results associated with Example 4.2, Second Stage

Unit	β_0^*	$\sum_{j\in E}\lambda_j^*(x_j, y_j)$	Input slack	Output slack	Unit being rated belongs to the
U1(2,2)	0	U3	0	8	Weak frontier
U2(4,2)	1	U3	0	8	Interior
U3(2,10)	0	U3	0	0	Strong frontier
U4(4,14)	0	U4	0	0	Strong frontier
U5(8,14)	2	U4	0	0	Weak frontier
U6(12,12)	4.5	0.5(U3 + U4) = (3,12)	0	0	Interior

Our second stage analysis shows, in this case, that the first stage projections of units U1 and U2 onto U1 are not strongly efficient, simply because U1 is not a strongly efficient point but a weak one.[22] The presence of a non-zero output slack associated to U1 reveals its nature (see second row of Table 2.5). Moreover, we are now able to classify each point and locate it on the strong frontier, on the weak part of the frontier or on the interior of the production possibility set (see last column of Table 2.5), without the help of any graphical display, which, by the way, are unavailable for any input-output space with dimensions greater than 3.

The BCC output-oriented model admits a similar treatment. Its original formulation follows.

$$
\begin{aligned}
BCC_{EFF}^{OO}(x_0, y_0) = Max \quad & \Phi_0 \\
s.t. \quad & \\
& \sum_{j\in E}\lambda_j x_{ij} \le x_{i0}, \quad i = 1, \ldots, m \\
& \sum_{j\in E}\lambda_j y_{rj} \ge \Phi_0 y_{r0}, \quad r = 1, \ldots, s \\
& \sum_{j\in E}\lambda_j = 1, \quad \lambda_j \ge 0, \quad j = 1, \ldots, n
\end{aligned}
\tag{2.13}
$$

The last model is a DEA efficiency model, and Φ_0^* is known as the output efficiency score of unit (x_0, y_0). Being aware that Φ_0^* is a quantity greater than or equal to 1, making the change of variable $\Phi_0 = 1 + \beta_0$ we just move from an efficiency model to an inefficiency model. Since maximizing Φ_0 is equivalent to maximizing $\beta_0 + 1$, the last model can be reformulated as the next *DDF* model,

[22] As early as in 1979, Charnes, Cooper and Rhodes were aware of this problem in relation to the CCR model (Charnes et al. 1978 and 1979). They published a mathematical solution to it some years later (Charnes and Cooper, 1984), based on the seminal paper of Charnes (1952). The functioning of the BCC model is completely similar.

provided we add slack variables to the restrictions in the same way as we did in model (2.12).

$$
\begin{aligned}
BCC^{OO}_{INEFF}(x_0, y_0; 0_m, y_0) = Max \quad & \beta_0 \\
s.t. \quad & \sum_{j\in E} \lambda_j x_{ij} + s^-_{i0} = x_{i0}, \quad i = 1, \ldots, m \\
& \sum_{j\in E} \lambda_j y_{rj} - s^+_{r0} = y_{r0} + \beta_0 y_{r0}, \quad r = 1, \ldots, s \\
& \sum_{j\in E} \lambda_j = 1, \\
& \lambda_j \geq 0, \quad j = 1, \ldots, n \\
& s^-_{i0}, s^+_{r0} \geq 0, \quad i = 1, \ldots, m,\ r = 1, \ldots, s
\end{aligned}
\tag{2.14}
$$

It is obvious that $\beta^*_0 \geq 0$ because $\Phi^*_0 \geq 1$. Since the directional vector $(0_m, y_0)$ is unable to change inputs, it is output-oriented. Moreover, since the change in outputs is guided by y_0 we pursue a proportional—or radial—augmentation of outputs. It is well established that this model does not necessarily project all the inefficient points onto the strongly efficient frontier, or, in other words, depending on the sample of units being rated, BCC^{OO}_{INEFF} can be a weak *DDF*. By applying the two stage procedure developed in Sect. 2.3 it is easy to end up with a comprehensive strong *DDF*, as well as with the corresponding derived strong inefficiency scores. Even in the simplest one input-one output space, the second stage could be needed, just as for the input-oriented version.

The next three sections are devoted to the three common economic inefficiency measures: cost inefficiency, revenue inefficiency and profit inefficiency. The choice depends on the way firms solve technical inefficiencies, by either reducing inputs, or expanding outputs, or both. Let us revise each of the three possibilities. In any of them we need to know the corresponding set of market prices, i.e., q_i, the unitary cost of input $i, i = 1, \ldots, m$, for the first case, p_r, the unitary price of output $r, r = 1, \ldots, s$, for the second case, or both, for the third case. To simplify notation, we denote the m-vector of input market prices by q, and the s-vector of output market prices by p.

2.5 Evaluating and Decomposing Cost Inefficiency Through the Associated *RDDF* Measure

Since, according to Sect. 2.2, $RDDF^{M,\Pi^M}$ has exactly the same behavior as M, in terms of detecting exactly the same technical inefficiency while keeping the same projection for each of the units under scrutiny, we can use the well-known Fenchel–Mahler inequality, developed for directional distance functions by Chambers et al.

(1998), and applied it to the *RDDF*.[23] We will assume in this section that M is an input-oriented model, i.e., for each unit being rated we are interested in reducing inputs as much as possible while maintaining the output levels. The strategy used for reducing inputs is determined by model M. As a direct consequence of Proposition 2.1 we can enunciate the next result.

Corollary 1.3 *If M is an input-oriented model, then its reverse-*DDF *is also an input-oriented model.*

Given an m-vector of unitary input costs, q, as well as a fix level of outputs, y_0, and assuming that T represents the production possibility set, the *cost function* is defined as

$$C(y_0, q) = \inf\{qx : (x, y_0) \in T\}. \tag{2.15}$$

In general, if T satisfies certain mathematical conditions, the infimum is reachable and we can switch from infimum to minimum (see Ray 2004). In particular, this happens for a DEA technology under VRS.

In order to evaluate $C(y_0, q)$ we assume, as before, that T is generated through a finite set of points $\{(x_j, y_j), j = 1, \ldots, n, x_j \geq 0, y_j \geq 0\}$. In this case we need to solve the next linear program.

$$\begin{aligned} C(y_0, q) = \underset{\lambda, x}{Min} \quad & \sum_{i=1}^{m} q_i x_i \\ s.t. \quad & \sum_{j=1}^{n} \lambda_j x_{ij} \leq x_i, \quad i = 1, \ldots, m \\ & \sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{r0}, \quad r = 1, \ldots, s \\ & \sum_{j=1}^{n} \lambda_j = 1, \\ & \lambda_j \geq 0, \quad j = 1, \ldots, n \\ & x_i \geq 0, \quad i = 1, \ldots, m \end{aligned} \tag{2.16}$$

Before considering the cost inefficiency decomposition through the corresponding Fenchel–Mahler inequality associated to the *RDDF* associated to model M, we need to define the cost inefficiency at point (x_0, y_0), after introducing the concept of cost deviation.

The *cost deviation* at point (x_0, y_0) is simply the difference between the cost at that point and the cost function—or minimum cost—at market input-prices q:

[23]Since model M is given, its associated *RDDF* can be either a weak *DDF* or a strong *DDF*, depending on the nature of M.

$$CD(x_0, y_0) := qx_0 - C(y_0, q) = \sum_{i=1}^{m} q_i x_{i0} - C(y_0, q). \tag{2.17}$$

For the sake of brevity, we write CD_0 for $CD(x_0, y_0)$.

Regarding the *DDF* and its existing dual relationship, we want to point out that it is possible to relate a term of normalized cost deviation, called cost inefficiency,CI_0, with the technical inefficiency detected by a given input-oriented DDF through the next inequality:

$$CI_0 := \frac{CD_0}{qg_{0x}} \geq \vec{D}(x_0, y_0; g_{0x}, 0). \tag{2.18}$$

Expression (2.18) is precisely the aforementioned Fenchel-Mahler inequality valid for any input-oriented *DDF* that allows cost inefficiency to be decomposed into the sum of technical inefficiency and an additive residual term identified as allocative inefficiency.

Next, resorting to the Fenchel-Mahler inequality associated to $RDDF^{M,\Pi^M}$ we get the next inequality,

$$CI_0^{M,\Pi^M} := \frac{CD_0^{M,\Pi^M}}{qg_{0x}^{M,\Pi^M}} \geq \tau_0^M. \tag{2.19}$$

The left hand-side term or cost inefficiency, CI_0^{M,Π^M}, satisfies a desirable index number property: it is homogeneous of degree 0 in prices, which makes CI_0^{M,Π^M} invariant to the currency units of the market input prices. As first pointed out by Nerlove (1965), CD_0^{M,Π^M}, the numerator of CI_0^{M,Π^M}, is homogeneous of degree 1 in prices and, consequently, the cost deviation cannot be considered as an appropriate economic measure. Going back to the last inequality and defining the *allocative inefficiency* as the corresponding additive residual, we get the next equality:

$$CI_0^{M,\Pi^M} = \tau_0^M + AI_0^{M,\Pi^M}. \tag{2.20}$$

In words, at point (x_0, y_0) cost inefficiency is decomposed into the sum of technical inefficiency and allocative inefficiency.

Example 5.1 The Input-Oriented Additive Model with Single Projections

Let us consider again model (2.3), known as the additive model (Charnes et al. 1985). It is a particular case of the weighted additive model (Lovell and Pastor 1995), where the objective function is simply a non-negative weighted sum of all the input and output slacks, with at least one positive weight. The additive model is a weighted additive model with all the weights equal to 1. Its input-oriented version

is obtained by setting the weights attached to the output slacks equal to 0 in the objective function. Its formulation is as follows.

$$
\begin{aligned}
Add_{IO}(x_0, y_0) = Max_{s^-,\lambda} \quad & \sum_{i=1}^{m} s_{i0}^- \\
s.t. & \\
& \sum_{j\in E} \lambda_j x_{ij} = x_{i0} - s_{i0}^-, \quad i = 1, \ldots, m \\
& \sum_{j\in E} \lambda_j y_{rj} = y_{r0} + s_{r0}^+, \quad r = 1, \ldots, s \\
& \sum_{j\in E} \lambda_j = 1, \\
& \lambda_j \geq 0, \quad j = 1, \ldots, n \\
& s_{i0}^- \geq 0, \quad i = 1, \ldots, m
\end{aligned}
\tag{2.21}
$$

Let us consider a sample of three units to be rated, defined as (4,7;1), (5,5;1) and (13,7;1). They belong to the "two input—one output" space. Since all the units have the same level of output we may represent them on the 2-dimensional input-plane. Moreover, focusing on the two inputs it is easy to realize that the first two units are efficient while the third one is clearly inefficient, because it is dominated by any of the other two efficient units.

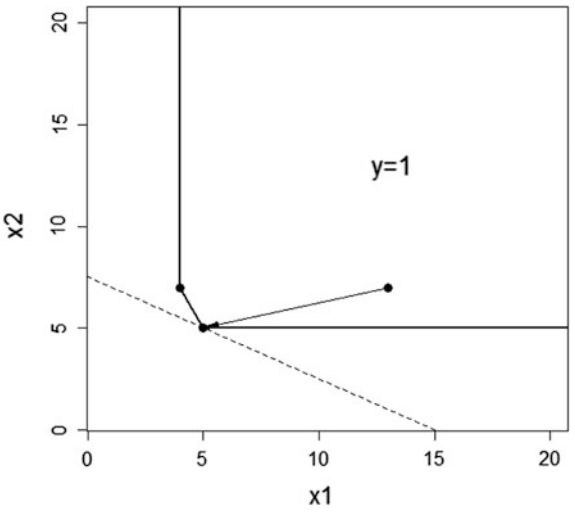


Fig. 2.2 Figure associated with Example 5.1

The L_1 projection of (13,7) onto (4,7) follows the L_1-path that connects both points, and whose components are (−9,0). Hence, the length of this L_1-path is 9 (=9 + 0). Alternatively, the L_1 projection of the inefficient unit onto (5,5) is given through the vector (−8,−2), which corresponds to a L_1-path of length 10 (=8 + 2). The maximum length is 10, which means that program (2.21) identifies (5,5) as the unique optimal projection[24] of unit (13,7), with an associated technical inefficiency, τ_0^M equal to 10. Since the projections are unique for all the three units in our example, we may write g_{0x}^M for g_{0x}^{M,Π^M}, as well as CI_0^M for CI_0^{M,Π^M} and AI_0^M for AI_0^{M,Π^M}.

Let us further assume that the unitary market input-prices are 1, for the first input, and 2, for the second. The minimum cost evaluated through program (2.16) tells us that $C(y_0,(q_1,q_2)) = C(1,(1,2)) = 15$, and that it is achieved at $\left(x_1^M, x_2^M\right) = (5,5)$. In Fig. 2.2 we have drawn the iso-cost line $x_1 + 2x_2 = 15$, which indeed passes through point (5,5). Now, since the obtained M-projection of point (x_0, y_0) = (13,7;1) is the strongly efficient point (5,5;1) and τ_0^M = 10, we get $g_{0x}^M = \left(\frac{8}{10}, \frac{2}{10}\right)$. In this example, $RDDF^M$ is perfectly defined knowing expression (2.2), that is, knowing that the directional vector at the only inefficient point (13,7;1) is $g_0^M = \left(\frac{8}{10}, \frac{2}{10}; 0\right)$. Evaluating the cost deviation at the unique inefficient point we obtain

$$CD(x_0,y_0) := \sum_{i=1}^{m} q_i x_{i0} - C(y_0,q) = (1,2)\cdot(13,7) - 15 = 13 + 14 - 15 = 12.$$

Consequently, the value of CI_0 is $\frac{CD_0}{qg_{0x}^M} = \frac{12}{1\cdot\frac{8}{10}+2\cdot\frac{2}{10}} = \frac{12\cdot 10}{12} = 10$, which equals τ_0^M = 10. Hence, $AI_0^M = CI_0^M - \tau_0^M = 10 - 10 = 0$.

The projection (5,5) of the inefficient unit (13,7) is, as said before, where the minimum cost is achieved for the considered market input-prices. Obviously this cost minimizing point has a cost deviation of 0, and also a cost inefficiency of 0, since $CI(5,5) = \frac{CD(5,5)}{(1,2)\cdot(1,1)} = \frac{0}{3} = 0$. Moreover, since (5,5) is efficient, its technical inefficiency is 0. Hence, its allocative inefficiency is also 0.

The previous result shows that point (13,7;1) is projected onto the minimum cost point (5,5;1), and has also an allocative inefficiency equal to 0. The next question springs to mind: is there any relationship between the allocative inefficiency of point (x_0, y_0) and the cost deviation of its efficient projection? The next proposition shows that the suggested relationship exists, and proposes an alternative way for evaluating the cost allocative inefficiency of point (x_0, y_0) based on its projection.

[24]Although in this simple example the projection is unique, it is straightforward to devise alternative easy examples for the input-oriented additive model where an inefficient unit may have two, or more, different projections. Our example gives rise to a single-value additive model, as opposed to a multiple-value additive model.

Proposition 2 *Let* $\left(x_0^M, y_0^M\right)$ *denote the projection of* (x_0, y_0) *obtained through model M. Let us assume that* $y_0^M = y_0$. *Then, the cost allocative inefficiency associated to point* (x_0, y_0) *and obtained through* $RDDF^{M,\Pi^M}$ *is*

$$AI_0^{M,\Pi^M} = \frac{CD^{M,\Pi^M}\left(x_0^M, y_0^M\right)}{q \cdot g_{0x}^{M,\Pi^M}} = \frac{\sum_{i=1}^m q_i x_{0i}^M - C(y_0^M, q)}{q \cdot g_{0x}^{M,\Pi^M}}. \tag{2.22}$$

In particular, $AI_0^{M,\Pi^M} = 0$ *if, and only if,* $CD\left(x_0^M, y_0^M\right) = 0$.

Proof If $(x_0, y_0) = \left(x_0^M, y_0^M\right)$, then $\tau_0^M = 0$, and (2.22) is a direct consequence of (2.20). Consequently, let us assume that (x_0, y_0) is an inefficient point. According to equalities (2.19) and (2.20), $AI_0^{M,\Pi^M} = CI_0^{M,\Pi^M} - \tau_0^M = \frac{CD_0}{q g_{0x}^{M,\Pi^M}} - \tau_0^M = \frac{CD_0 - \tau_0^M q g_{0x}^{M,\Pi^M}}{q g_{0x}^{M,\Pi^M}}$. Hence, taking into account expression (2.22), all we need to prove is that $CD\left(x_0^M, y_0^M\right) = CD_0 - \tau_0^M q g_{0x}^{M,\Pi^M}$, or, equivalently, that $\sum_{i=1}^m q_i x_{i0}^M - C(y_0^M, q) = \left\{\sum_{i=1}^m q_i x_{i0} - C(y_0, q)\right\} - \tau_0^M q g_{0x}^{M,\Pi^M}$. According to expression (2.2) the equality $q x_0^M = q\left(x_0 - \tau_0^M g_{0x}^{M,\Pi^M}\right)$ holds. Therefore, the previous expression can be reduced to $C(y_0^M, q) = C(y_0, q)$ which trivially holds because we are assuming that $y_0^M = y_0$. ■

Proposition 2 shows that the factor to be used for normalizing the cost deviation associated to the efficient projection so as to obtain the allocative inefficiency of the inefficient point is exactly the normalization factor associated to the inefficient point. Moreover, it seems an acceptable and intuitive property that, when the projection is a cost minimizing point, the allocative inefficiency associated to the point being rated is 0.

Example 5.2 The Input-Oriented Additive Model with Multiple Projections

Let us consider again, in the two input—one output space, a sample of three units to be rated, defined as (3,7;1), (5,5;1) and (13,7;1). In comparison to Example 5.1 we have only slightly changed the first unit, from (4,7;1) to (3,7;1). This change does not affect the efficiency status of the three units, but does affect the L_1-distance from (13,7) to the first efficient unit (3,7), which has increased and takes exactly the same value, 10, as the distance from (13,7) to the second efficient unit, (5,5). Now the two efficient units are optimal projections for the unique inefficient unit, or, in other words, (13,7) has multiple optimal projections. The preferable option can be a function of a second criterion. For instance, if the market input-prices are again q = (1,2) we might prefer to select the projection that generates a lower allocative inefficiency, or, according to Proposition 2, the projection were the lowest cost deviation is achieved. Since, the new considered efficient point (3,7) has an

input-cost of $1 \cdot 3 + 2 \cdot 7 = 17$, higher than the already evaluated input-cost of point (5,5), which equals 15 and corresponds to the value of the cost function—see Example 4.1—, then the cost deviation of the new obtained projection, $CD(3,7) = 17 - 15 = 2$, is greater than the cost deviation of the old one, $CD(5,5) = 0$. Hence, our final choice is to select point (5,5) as our preferred projection, according to our aforementioned second criterion, because its allocative inefficiency is the most convenient one.

In this simple example we have directly considered the two possible alternative optimal projections as points of our sample. When solving a real life problem, with many units and a higher number of inputs and outputs, a specific search for identifying at each inefficient unit of the sample possible alternative optimal projections needs to be developed. This task is accomplished in Appendix 1.

2.6 Evaluating and Decomposing Revenue Inefficiency

Given an s-vector of unitary output prices, $p \geq 0_s$, and being T the production possibility set, the *revenue function* is defined for a fix level of inputs, x_0, as follows.

$$R(x_0, p) = \sup\{py : (x_0, y) \in T\}. \tag{2.23}$$

In the case of T being a DEA technology, the supremum in (2.23) may be equivalently changed by maximum.

In order to evaluate $R(x_0, p)$ we assume that the set $\{p_r, r = 1, \ldots, s\}$ of non-negative output prices is known and that, within a DEA framework, T is generated through a finite set of n points $\{(x_j, y_j), j = 1, \ldots, n, x_j \geq 0, y_j \geq 0\}$. In this case we only need to solve the next linear program:

$$\begin{aligned} R(x_0, p) = \underset{\lambda, y}{Max} \quad & \sum_{r=1}^{s} p_r y_r \\ s.t. \quad & \\ & \sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{i0}, \quad i = 1, \ldots, m \\ & \sum_{j=1}^{n} \lambda_j y_{rj} \geq y_r, \quad r = 1, \ldots, s \\ & \sum_{j=1}^{n} \lambda_j = 1, \\ & \lambda_j \geq 0, \quad j = 1, \ldots, n \\ & y_r \geq 0, \quad r = 1, \ldots, s \end{aligned} \tag{2.24}$$

Before considering the revenue inefficiency decomposition associated to $RDDF^{M,\Pi^M}$ through the corresponding Fenchel–Mahler inequality, we need to define the revenue deviation at point (x_0, y_0).

The *revenue deviation* at point (x_0, y_0) is simply the difference between the revenue function and the revenue at that point, given market output-prices p:

$$RD(x_0, y_0) := R(x_0, p) - py_0. \tag{2.25}$$

For the sake of brevity, we write RD_0 for $RD(x_0, y_0)$.

As for the *DDF* and thanks to its dual relationships, it is possible to link, at point (x_0, y_0), a normalized term of revenue deviation, called *revenue inefficiency*, with the optimal value of any output-oriented directional distance function, as follows:

$$RI_0 := \frac{RD_0}{pg_{0y}} \geq \vec{D}\left(x_0, y_0; 0, g_{0y}\right). \tag{2.26}$$

Expression (2.26) is the Fenchel-Mahler inequality associated with any *DDF* that allows decomposing revenue inefficiency into the sum of technical inefficiency and allocative inefficiency.

In our particular case, and for the $RDDF^{M,\Pi^M}$, we get the inequality

$$RI_0^{M,\Pi^M} := \frac{RD_0^{M,\Pi^M}}{pg_{0y}^{M,\Pi^M}} \geq \tau_0^{M,\Pi^M}. \tag{2.27}$$

The left hand-side term is the normalized revenue deviation, also called *revenue inefficiency,* RI_0, which, as CI_0, satisfies that it is homogeneous of degree 0 in prices, which makes RI_0 invariant to the currency units associated to the market output prices. Going back to the last inequality and defining the revenue allocative inefficiency as the corresponding residual, we get the next equality:

$$RI_0^{M,\Pi^M} = \tau_0^{M,\Pi^M} + AI_0^{M,\Pi^M}. \tag{2.28}$$

In words, at point (x_0, y_0), and thanks to the associated $RDDF^{M,\Pi^M}$, normalized revenue inefficiency is decomposed into the sum of technical inefficiency and allocative inefficiency.

Example 6.1 The Output-Oriented Additive Model with Single Projections

Let us now consider as model M the output-oriented version of the additive model, whose objective function only includes output slacks. Its formulation follows.

$$
\begin{aligned}
Add_O(x_0, y_0) = Max_{s^+,\lambda} \quad & \sum_{r=1}^{s} s_{r0}^{+} \\
& s.t. \\
& \sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{i0}, \quad i = 1, \ldots, m \\
& \sum_{j=1}^{n} \lambda_j y_{rj} = y_{r0} + s_{r0}^{+}, \quad r = 1, \ldots, s \\
& \sum_{j=1}^{n} \lambda_j = 1, \\
& \lambda_j \geq 0, \quad j = 1, \ldots, n \\
& s_{r0}^{+} \geq 0, \quad r = 1, \ldots, s
\end{aligned}
\tag{2.29}
$$

Let us consider now in the "one input—two output space" the next sample of units to be rated: $\{(1;3,5), (1;5,2); (1;2,1)\}$. Since all the units have the same level of input we may represent them on the 2-dimensional output plane. Moreover, focusing on the two outputs it is easy to realize that the first two units are efficient while the third one is inefficient because it is dominated by any of the two efficient units.

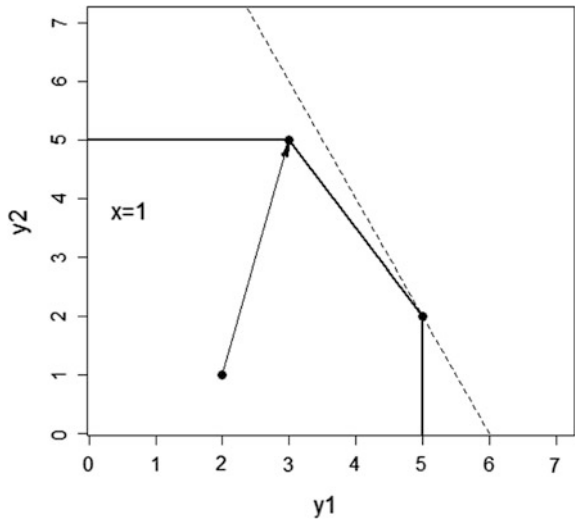


Fig. 2.3 Figure associated with Example 6.1

The L_1 projection of (2,1) onto (3,5) follows the L_1-path that connects both points, and whose components are (1,4). Hence, the length of this L_1-path is 5 (=1 +4). Alternatively, the L_1 projection of the inefficient unit onto (5,2) is given through the vector (3,1) which corresponds to a L_1-path of length 4 (=3+1). The maximum length is 5, which means that program (2.29) will identify (3,5) as the unique optimal projection with an associated technical inefficiency, τ_0^M, as given by the optimal value of the objective function, equal to 5. Let us assume that the market output-prices are given as p = (2,1). The maximum revenue—or revenue function—, as evaluated through program (2.24), tell us that $R(1,(2,1)) = 12$, and that it is achieved at point (5,2), as shown in Fig. 2.3, where we have drawn the iso-revenue line $2y_1 + y_2 = 12$. Now, since the obtained M-projection of point (x_0, y_0) = (1;2,1) is the efficient point (1;3,5) and τ_0^M = 5, we get $g_{y0} = \left(\frac{1}{5}, \frac{4}{5}\right)$. Evaluating the revenue deviation at the inefficient point we obtain $RD(x_0, y_0) := R(x_0, p) - \sum_{r=1}^{s} p_r y_{0r} = 12 - (2,1) \cdot (2,1) = 12 - 5 = 7$. Consequently, the value of RI_0 is $\frac{RD_0}{pg_{0y}^M} = \frac{7}{2 \cdot \frac{1}{5} + 1 \cdot \frac{4}{5}} = \frac{7 \cdot 5}{6} = \frac{35}{6}$, which is greater than $\tau_0^M = 5$. Hence, $AI_0^M = RI_0^M - \tau_0^M = \frac{35}{6} - 5 = \frac{5}{6}$.

It would be interesting, as we did when evaluating cost allocative inefficiency, to connect the revenue deviation of the corresponding efficient projection with the revenue allocative inefficiency of the point being rated. The next proposition gives us the clue.

Proposition 3 *Let* $\left(x_0^M, y_0^M\right)$ *denote the projection of* (x_0, y_0) *obtained through model M. Let us assume that* $x_0^M = x_0$. *Then, the revenue allocative inefficiency associated to point* (x_0, y_0) *can be obtained as*

$$AI_0^{M,\Pi^M} = \frac{RD_0^{M,\Pi^M}\left(x_0^M, y_0^M\right)}{p \cdot g_{0y}^{M,\Pi^M}} = \frac{R(x_0^M, p) - \sum_{r=1}^{s} p_r y_{0r}^M}{p \cdot g_{0y}^{M,\Pi^M}}. \tag{2.30}$$

In particular, $AI_0^{M,\Pi^M} = 0$ *if, and only if,* $RD_0^{M,\Pi^M}\left(x_0^M, y_0^M\right) = 0$.

Proof The proof is similar to the proof of Proposition 2 and is left to the reader. ■

2.7 Evaluating and Decomposing Profit Inefficiency

The profit function requires that both market input costs and market output revenues are specified, by knowing the corresponding market unitary prices. As usual, let us denote by $q \geq 0_m$ the market input-prices and by $p \geq 0_s$ the market output-prices. The *profit function* is defined as follows.

$$\Pi(q,p) = \sup\{py - qx : (x,y) \in T\}. \tag{2.31}$$

Under the hypothesis of working with a DEA production possibility set, the supremum in (2.31) is reachable and we switch from supremum to maximum.

Within a DEA framework, the linear program to be solved in order to calculate the profit function is the next one.

$$\begin{aligned}
\Pi(q,p) = \underset{\lambda,x,y}{Max} & \sum_{r=1}^{s} p_r y_r - \sum_{i=1}^{m} q_i x_i \\
& s.t. \\
& \sum_{j=1}^{n} \lambda_j x_{ij} \le x_i, \quad i = 1,\ldots,m \\
& \sum_{j=1}^{n} \lambda_j y_{rj} \ge y_r, \quad r = 1,\ldots,s \\
& \sum_{j=1}^{n} \lambda_j = 1, \\
& \lambda_j \ge 0, \quad j = 1,\ldots,n \\
& x_i \ge 0, y_r \ge 0 \quad i = 1,\ldots,m,\ r = 1,\ldots,s
\end{aligned} \tag{2.32}$$

As usual, (2.32) is a VRS model. In fact, considering a CRS model could be seen as meaningless from an entrepreneur's point of view when the aim is to measure profit inefficiency, because the CRS assumption implies always either unbounded profit or zero maximal profit. Nevertheless, it would be possible to use another alternative hypothesis on the production possibility set as NIRS (Non-Increasing Returns to Scale), where the constraint $\sum_{j=1}^{n} \lambda_j = 1$ in (2.32) would be substituted by $\sum_{j=1}^{n} \lambda_j \le 1$.

Before considering the profit inefficiency decomposition through the corresponding Fenchel–Mahler inequality we need to define the profit deviation at point (x_0, y_0).

The *profit deviation* at point (x_0, y_0) is simply the deviation between the profit function and the profit at that point, given market prices (q, p):

$$\Pi D(x_0,y_0) := \Pi(q,p) - (py_0 - qx_0). \tag{2.33}$$

For the sake of brevity, we write ΠD_0 for $\Pi D(x_0, y_0)$.

Additionally, it is possible to relate a normalized term of profit deviation, called *profit inefficiency*, with the inefficiency detected by the directional distance function as follows:

$$\frac{\Pi D_0}{pg_{0y} + qg_{0x}} \geq \vec{D}\left(x_0, y_0; g_{0x}, g_{0y}\right). \tag{2.34}$$

Resorting to the Fenchel–Mahler inequality associated to $RDDF^M$ we get the inequality

$$\Pi I_0^{M,\Pi^M} := \frac{\Pi D_0^{M,\Pi^M}}{pg_{0y}^{M,\Pi^M} + qg_{0x}^{M,\Pi^M}} \geq \tau_0^M. \tag{2.35}$$

The left hand-side term is the profit inefficiency, ΠI_0, which, as CI_0 and RI_0, satisfies that it is homogeneous of degree 0 in prices, which makes ΠI_0 invariant to the currency units for the market output and input prices. Going back to the last inequality and defining the profit allocative inefficiency as the corresponding residual, we get the next equality:

$$\Pi I_0^{M,\Pi^M} = \tau_0^M + AI_0^{M,\Pi^M}. \tag{2.36}$$

In other words, at point (x_0, y_0), profit inefficiency is decomposed into the sum of technical inefficiency and allocative inefficiency.

Example 7.1 The Additive Model with Single Projections

Let us consider model M as the additive model, formulated before and identified as model (2.3). Let us further consider in the "one input—one output space" the next sample of units to be rated: $\{(2;2), (6;10); (8;6)\}$. We represent them directly on a plane. Moreover, it is easy to realize that the first two units are efficient while the third one is not, because it is clearly dominated by (6;10).

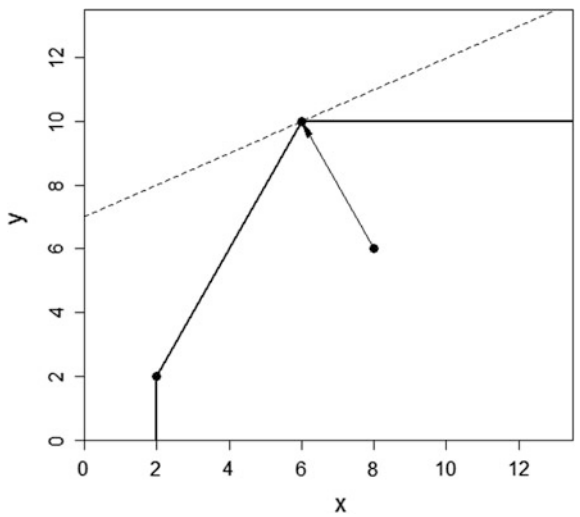


Fig. 2.4 Figure associated with Example 7.1

Graphically, there are several candidates as projection of the inefficient unit (8;6): the point (6;10) or certain convex linear combinations of the two efficient points that dominates point (8;6), such as point (4;6), the middle point in between (2;2) and (6;10). Additive model M selects the one that is as far as possible, using as measure the L_1-distance. The projection is point (6;10) with a L_1-path length equal to 6 (= (8 − 6) + 10 − 6). This length is exactly the technical inefficiency, τ_0^M, detected by linear program (2.3). Let us assume that the market prices are given as $(q,p) = (2, 4)$. The maximum profit as evaluated through program (2.32) tell us that $\Pi(2,4) = 40 - 12 = 28$, and that it is achieved at point (6;10), as shown in Fig. 2.4, where we have drawn the iso-profit line $4y - 2x = 28$. Now, since the obtained M-projection of point $(x_0, y_0) = (8;6)$ is the efficient point (6;10) and $\tau_0^M = 6$, we get $g_0 = \left(\frac{2}{6}, \frac{4}{6}\right) = \left(\frac{1}{3}, \frac{2}{3}\right)$. Evaluating the profit deviation at the inefficient point we obtain $\Pi D(x_0, y_0) := \Pi(2,4) - \left(\sum_{r=1}^{s} p_r y_{r0} - \sum_{i=1}^{m} q_i x_{i0}\right) = 28 - (4 \cdot 6 - 2 \cdot 8) = 28 - 8 = 20$. Consequently, the value of ΠI_0 is $\frac{\Pi D_0}{q g_{0x}^M + p g_{0y}^M} = \frac{20}{2 \cdot \frac{1}{3} + 4 \cdot \frac{2}{3}} = \frac{20}{\frac{10}{3}} = 6$, which is equal to $\tau_0^M = 6$. Hence, $AI_0^M = \Pi I_0^M - \tau_0^M = 6 - 6 = 0$.

Once again, it would be interesting if we could relate the profit deviation of the efficient projection with the allocative inefficiency of the point being rated. The next proposition gives us the answer.

Proposition 4 *Let* $\left(x_0^M, y_0^M\right)$ *denote the projection of* (x_0, y_0) *obtained through model M. Then, the allocative inefficiency associated to point* (x_0, y_0) *can be obtained as*

$$AI_0^{M,\Pi^M} = \frac{\Pi D_0^{M,\Pi^M}\left(x_0^M, y_0^M\right)}{q \cdot g_{0x}^{M,\Pi^M} + p \cdot g_{0y}^{M,\Pi^M}} = \frac{\Pi(q,p) - \left(\sum_{r=1}^{s} p_r y_{r0}^M - \sum_{i=1}^{m} q_i x_{i0}^M\right)}{q \cdot g_{0x}^{M,\Pi^M} + p \cdot g_{0y}^{M,\Pi^M}}. \quad (2.37)$$

In particular, $AI_0^{M,\Pi^M} = 0$ *if, and only if,* $\Pi D_0^{M,\Pi^M}\left(x_0^M, y_0^M\right) = 0$.

The proof is similar to the proof of Proposition 2 and is left again to the reader. The key of the proof is that profit at the efficient projection is equal to profit at the inefficient point plus the technical inefficiency times the normalization factor of the inefficient point.

2.8 Identifying, for Each Inefficient Unit, a Projection that Minimizes Its *RDDF* Profit Inefficiency

For a specific inefficient unit, (x_0, y_0), the considered DEA single-value model M generates a specific efficient projection, $\left(x_0^M, y_0^M\right)$. Resorting to $RDDF^M$, we have been able to measure and decompose its economic inefficiency. Let us focus our

attention on profit inefficiency, knowing that a completely similar treatment can be developed for cost or revenue inefficiency. The question that we want to tackle is the following: is it possible to identify a different projection with better—or lower—profit inefficiency? Let us refer to this new projection as (x_0^*, y_0^*). Since our additional aim is to maintain the introduced profit decomposition through the *RDDF*, we will not accept the possibility of increasing some inputs and decreasing some outputs, as Zofío et al. (2013) did.

Being our aim to reduce profit inefficiency as much as possible, we have devised the following strategy. First of all, let us observe that according to expression (2.35) profit inefficiency equals $\frac{\Pi D_0^M}{pg_{0y}^M + qg_{0x}^M}$. Hence, it is a ratio whose numerator, the profit deviation of point (x_0, y_0), is fixed, and, consequently, if we want to reduce profit inefficiency, the only action we can take is to enlarge its denominator. Therefore, what we would like to do is to search for an efficient projection (x_0^*, y_0^*) that maximizes this denominator. According to expression (2.2), and for any inefficient point (x_0, y_0), we know that[25]

$$q \cdot g_{0x}^* + p \cdot g_{0y}^* = \frac{1}{\tau_0^{*M}} \left[q \cdot \left(x_0 - x_0^* \right) + p \cdot \left(y_0^* - y_0 \right) \right]. \tag{2.38}$$

Hence, if we maximize $q \cdot \left(x_0 - x_0^* \right) + p \cdot \left(y_0^* - y_0 \right)$ we are maximizing $\tau_0^{*M} \left(q \cdot g_{0x}^* + p \cdot g_{0y}^* \right)$. This is not exactly what we want to do, but it is a useful proxy that will help us to achieve our goal. In fact, assuming that $\tau_0^{*M} \leq \tau_0^M$,[26] we obtain the next chain of inequalities: $q \cdot g_{0x}^M + p \cdot g_{0y}^M = \frac{1}{\tau_0^M} \left[q \cdot \left(x_0 - x_0^M \right) + p \cdot \left(y_0^M - y_0 \right) \right]$ $\leq \frac{1}{\tau_0^{*M}} \left[q \cdot \left(x_0 - x_0^M \right) + p \cdot \left(y_0^M - y_0 \right) \right] \leq \frac{1}{\tau_0^{*M}} \left[q \cdot \left(x_0 - x_0^* \right) + p \cdot \left(y_0^* - y_0 \right) \right] = q \cdot g_{0x}^*$ $+ p \cdot g_{0y}^*$, where the last inequality is true because (x_0^*, y_0^*) is the efficient point that dominates (x_0, y_0) and, at the same time, maximizes $q \cdot \left(x_0 - x_0^* \right) + p \cdot \left(y_0^* - y_0 \right)$.

Consequently, let us maximize expression $q \cdot \left(x_0 - x_0^* \right) + p \cdot \left(y_0^* - y_0 \right)$, or, equivalently, $\left(p \cdot y_0^* - q \cdot x_0^* \right) - \left(p \cdot y_0 - q \cdot x_0 \right)$. Since the last parenthesis is a fixed number, we simply need to maximize $\left(p \cdot y_0^* - q \cdot x_0^* \right)$, i.e., the profit achieved at the new efficient projection. Let us consider a linear program whose objective function

[25] The presence of the asterisk means that we are considering a new projection obtained through a specific optimization program and not through model *M*. Nonetheless, model *M* is used for determining the technical inefficiency associated to this new projection, which justifies the used notation.

[26] The assumption is valid for DEA models that reach their projections by maximizing a certain "distance", such as the weighted additive model.

maximizes profit at the efficient projection and whose restrictions guarantee that the obtained projection belongs to T. At this point it is worth noticing that a few years ago, the idea of getting a reference benchmark with the highest possible profit was already suggested by Zofio et al. (2013, page 263, Footnote 3).

$$
\begin{aligned}
Max_{s^-,s^+,\lambda} & \sum_{r=1}^{s} p_r\left(y_{r0}+s_{r0}^{+}\right)-\sum_{i=1}^{m} q_i\left(x_{i0}-s_{i0}^{-}\right) \\
& s.t. \\
& \sum_{j\in E}\lambda_j x_{ij}=x_{i0}-s_{i0}^{-}, \quad i=1,\ldots,m \\
& \sum_{j\in E}\lambda_j y_{rj}=y_{r0}+s_{r0}^{+}, \quad r=1,\ldots,s \\
& \sum_{j\in E}\lambda_j=1, \\
& \lambda_j\geq 0, \quad j\in E \\
& s_{i0}^{-}\geq 0, \quad i=1,\ldots,m \\
& s_{r0}^{+}\geq 0, \quad r=1,\ldots,s
\end{aligned}
\tag{2.39}
$$

Program (2.39) identifies $\left(x_0^*, y_0^*\right)$ as a strongly efficient point that maximizes $\left(p\cdot y_0^*-q\cdot x_0^*\right)$[27] and, at the same time, dominates (x_0, y_0). The used notation together with program (2.39) indicate that the last identified efficient point is valid for any non-oriented M model.[28] The differences will appear when we evaluate the technical inefficiency associated to that strongly efficient point through model M. For instance, if model M is the additive model, the associated τ_0^{*M} is the length of the L_1-path connecting (x_0, y_0) with $\left(x_0^*, y_0^*\right)$, and it is very likely that $\tau_0^{*M}<\tau_0^M$. Let us illustrate these findings with an easy numerical example.

Example 8.1 Maximizing the *RDDF* Profit Inefficiency of the Additive Model

Let us consider the next sample of points on the XY plane:$\{(2,2),(3,4),(5,7),(11,9),(8,3)\}$. Resorting to the additive model it is easy to check that all the points are strongly efficient except the last one. In fact, point (8,3) is dominated by point (3,4) as well as by point (5,7). Let us further assume that the market prices are $q = 1$ and $p = 11$. Figure 2.5 below shows the graphical representation of the sample of points as well as of the corresponding VRS efficient frontier. We have also drawn the line associated to maximum profit, $11y - x = 88$.

[27] In DEA literature it is known that maximum profit is achieved in at least a strongly efficient point.

[28] The other two possibilities are that model M is input-oriented or output-oriented. In the first case, program [39] needs to be adjusted just by deleting the output-slacks, and symmetrically for the output-oriented case.

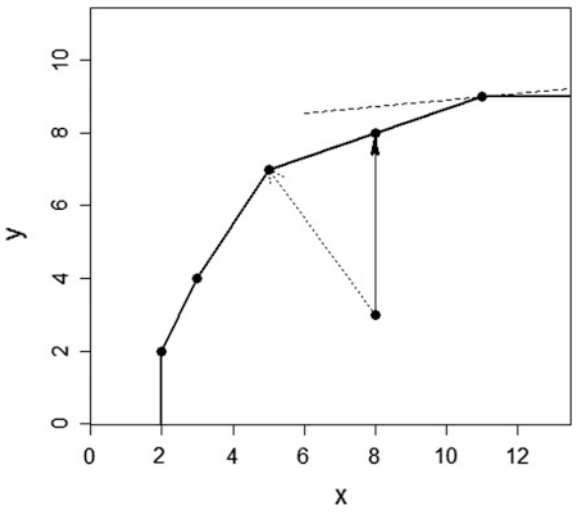


Fig. 2.5 Figure associated with Example 8.1

For the only inefficient point, $(x_0, y_0) = (8,3)$, we solve linear program (2.39) and obtain the profit maximizing projection $\left(x_0^*, y_0^*\right) = (8,8)$, that belongs to the facet defined by efficient points (5,7) and (11,9). The profit at point (8,8) is 80 (=11 · 8 − 1 · 8), the profit at point (8,3) is 25 (=33 − 8), while the optimal profit, $\Pi(1,2) = 88$, is achieved at point (11,9). Let us now consider that model M is the additive model. Clearly, the technical inefficiency associated to the new obtained projection equals 5 (length of the L_1-path connecting (8,3) with (8,8)). Moreover, $q \cdot g_{0x}^{*M} + p \cdot g_{0y}^{*M} = 1 \cdot 0 + 11 \cdot 5 = 55$. Consequently, according to expression (2.35), the normalized profit inefficiency at point (8,3) is $\frac{88-25}{55} = \frac{63}{55} = 1\frac{8}{55}$.

In order to evaluate the gain obtained by applying the new proposed strategy, let us compare the last result with the result derived directly by applying the additive model and evaluating the associated normalized profit inefficiency through the $RDDF^M$. Since the additive projection of point (8,3) is point (5,7), the corresponding τ_0^M equals 7(=(8 − 5) + (7 − 3)) which means that $\left(g_{0x}^M, g_{0y}^M\right) = \frac{1}{\tau_0^M}\left(x_0 - x_0^M, y_0^M - y_0\right) = \left(\frac{3}{7}, \frac{4}{7}\right)$, with a normalization factor value equal to $1 \cdot \frac{3}{7} + 11 \cdot \frac{4}{7} = \frac{47}{7}$, which gives a normalized profit inefficiency value equal to $\frac{63}{\frac{47}{7}} = \frac{63 \cdot 7}{47} = \frac{441}{47} = 9\frac{8}{47}$, clearly much bigger than $1\frac{8}{55}$.

2.9 Conclusions

The introduction, in Sect. 2.2, of the *RDDF* associated to DEA inefficiency model M, denoted as $RDDF^{M,\Pi^M}$, has allowed us, for the first time, to express any DEA model as a *DDF*. The key idea is that the *RDDF* maintains exactly the same

projections as the original DEA model. However, if the original DEA model has multiple projections for at least one inefficient unit, we are able to define as many *RDDFs* as combinations of single projections we can perform. For the case of the weighted additive model we have included, in Appendix 1, a new method for identifying different projections at each inefficient point. Which projection to choose will depend on secondary criteria that will guide our final selection.

Our new introduced tool is also relevant for transforming a weak *DDF* into a comprehensive measure or strong *DDF*. Moreover, any DEA inefficiency measure can also be transformed into its comprehensive version, by applying the *RDDF* technique twice. In particular, we have shown how to transform a radial model into a comprehensive one that is likely to lose its radial type of projection, as well as its one-sided orientation.

The first introduced (multiplicative) decomposition of economic efficiency, cost efficiency in particular, was due to Farrell, designed specifically for radial models. In the nineties Chambers, Chung and Färe proposed an additive decomposition of economic inefficiency, cost, revenue or profit, based on the *DDF*. So far, the subsequent proposed approaches for estimating and decomposing economic inefficiency have been all additive in nature and have emerged during the last lustrum. As explained in Sect. 2.1, the models that have captured the attention of the researchers in DEA were the weighted additive model, the output-oriented weighted additive model, and the two Russell oriented models. The new *RDDF* introduced in this chapter is responsible for defining the cost, revenue or profit inefficiency of any DEA inefficiency measure as well as their additive decomposition into its technical and allocative components. Hence the proposed solution constitutes a unified DEA approach that benefits from the known Fenchel–Mahler inequality established for *DDFs*.

Finally, two additional issues have been considered and solved. First, a linear programming procedure has been devised for identifying a new projection for each inefficient unit where profit inefficiency is minimized. And secondly, in Appendix 1, we have shown how to generate alternative optimal solutions in connection with additive type models.

Appendix 1

How to Search for Alternative Optimal Solutions When Using a Weighted Additive Model

The additive model has been introduced in Sect. 2.3, while the weighted additive model (Lovell and Pastor, 1995) has been described at the beginning of Example 4.1. Just as a reminder, the weighted additive model has the same set of restrictions as the additive model but differs in its objective function. In fact, while the objective function of the additive model is the sum of input-slacks and output-slacks, the

weighted additive model considers as objective function a weighted sum of input-slacks and output slacks, where the attached weights are all non-negative and at least one of them must be positive. Since any particular weight can be 0, the input-oriented or output-oriented weighted additive models are particular cases of weighted additive models. Moreover, weighted additive models are VRS models, because the convexity constraint is one of its restrictions. It is easy to consider CRS weighted additive models just by deleting from the set of restrictions the mentioned constraint, or even to consider non-increasing returns to scale (NIRS) or non-decreasing returns to scale (NDRS) additive models, by changing slightly the convexity constraint, transforming the equality into an inequality (≤ 1 for NIRS or ≥ 1 for NDRS). Here, as in the rest of the chapter, we will deal exclusively with VRS models, but it is not difficult to derive the corresponding conclusions for non-VRS weighted additive models.

Assume that, for a given inefficient unit (x_0, y_0) we have obtained a first optimal projection identified as $\left(x_0^{p1}, y_0^{p1}\right)$ through the weighted additive model. This projection is always a strongly efficient point. Now we want to search for the existence of alternative optimal solutions, that is, alternative optimal slack values. Let us denote as $w_i^-, i = 1, \ldots, m$ the weights associated to the input-slacks and as $w_r^+, r = 1, \ldots, s$ the weights associated to the output slacks in the objective function. Then, the optimal value of the objective function equals $\sum_{i=1}^m w_i^- \left(x_i^{p1} - x_{i0}\right) + \sum_{r=1}^s w_r^+ \left(y_r^{p1} - y_{r0}\right) = \sum_{i=1}^m w_i^- s_{i0}^{-*} + \sum_{r=1}^s w_r^+ s_{r0}^{+*}$, which is a fixed number, let us say v^*. Knowing v^*, we are able to generate as much as $2m + 2s$ optimal solutions through the procedure proposed next.

Procedure Consider the following linear program, which is equivalent to (2.3) except for the presence of non-negative weights in the objective function.

$$
\begin{aligned}
WAdd(x_0, y_0) = Max_{s^-, s^+, \lambda} \quad & \sum_{i=1}^m w_i^- s_{i0}^- + \sum_{r=1}^s w_r^+ s_{r0}^+ \\
s.t. \quad & \\
& \sum_{j=1}^n \lambda_j x_{ij} = x_{i0} - s_{i0}^-, \quad i = 1, \ldots, m \\
& \sum_{j=1}^n \lambda_j y_{rj} = y_{r0} + s_{r0}^+, \quad r = 1, \ldots, s \\
& \sum_{j=1}^n \lambda_j = 1, \\
& \lambda_j \geq 0, \quad j = 1, \ldots, n, \\
& s_{i0}^- \geq 0, \quad i = 1, \ldots, m, \quad s_{r0}^+ \geq 0, \quad r = 1, \ldots, s
\end{aligned}
\tag{2.40}
$$

As said before, and to simplify the notation, we write $v^* := WAdd(x_0, y_0)$.

In order to search for alternative optimal solutions, we further solve the next pair of linear programs for each input slack, $s_{k0}^-, k \in \{1,\ldots,m\}$ and each output slack, $s_{l0}^+, l \in \{1,\ldots,s\}$:

$$
\begin{aligned}
Max_{s^-,\lambda} \quad & s_{k0}^- (or\, s_{l0}^+) \\
& s.t. \\
& \sum_{j=1}^{n} \lambda_j x_{ij} = x_{i0} - s_{i0}^-, \quad i = 1,\ldots,m \\
& \sum_{j=1}^{n} \lambda_j y_{rj} = y_{r0} + s_{r0}^+, \quad r = 1,\ldots,s \\
& \sum_{j=1}^{n} \lambda_j = 1, \\
& \sum_{i=1}^{m} w_i^- s_{i0}^- + \sum_{r=1}^{s} w_r^+ s_{r0}^+ = v^* \\
& \lambda_j \geq 0, \quad j = 1,\ldots,n, \\
& s_{i0}^- \geq 0, \quad i = 1,\ldots,m, \;\; s_{r0}^+ \geq 0, \;\; r = 1,\ldots,s \\
& s_{i0}^- \geq 0, \quad i = 1,\ldots,m, \;\; s_{r0}^+ \geq 0, \;\; r = 1,\ldots,s,
\end{aligned}
\tag{2.42}
$$

and

$$
\begin{aligned}
Min_{s^-,\lambda} \quad & s_{k0}^- (or \quad s_{l0}^+) \\
& s.t. \\
& \sum_{j=1}^{n} \lambda_j x_{ij} = x_{i0} - s_{i0}^-, \quad i = 1,\ldots,m \\
& \sum_{j=1}^{n} \lambda_j y_{rj} = y_{r0} + s_{r0}^+, \quad r = 1,\ldots,s \\
& \sum_{j=1}^{n} \lambda_j = 1, \\
& \sum_{i=1}^{m} w_i^- s_{i0}^- + \sum_{r=1}^{s} w_r^+ s_{r0}^+ = v^* \\
& \lambda_j \geq 0, \quad j = 1,\ldots,n, \\
& s_{i0}^- \geq 0, \quad i = 1,\ldots,m, \;\; s_{r0}^+ \geq 0, \;\; r = 1,\ldots,s,
\end{aligned}
\tag{2.43}
$$

which means that we propose to solve $2m + 2s$ linear programs. Any of these programs searches for a possibly alternative optimal projection due to the addition of the last linear restriction $\sum_{i=1}^{m} w_i^- s_{i0}^- + \sum_{r=1}^{s} w_r^+ s_{r0}^+ = v^*$.

Example A.1.2 Searching for Alternative Optimal Solutions

Let us consider in the two input—one output space the next sample of units: U1 (1,6;3), U2(2,6;4), U3(6,2;4), U4(6,1;3) and U5(8,7;2). Let us work with the additive model, which is a weighted additive model where all weights equal 1. It is easy to check that the first four units are extreme strongly efficient and that U5 is inefficient. The projection we get resorting to Excel-Solver is point U3, with an optimal value equal to 9.

Now we start searching for alternative optimal solutions, with $v^* = 9$, by considering each of the maximizing models gathered in (2.42). For the objective function s_{15}^- we get U1 as a new alternative optimal solution, while for the objective function s_{25}^- we also get U4 as a new solution. Finally, for s_5^+ as the objective function we once more get U3 as solution. Going over to the minimizing models included in (2.43), and starting again with s_{15}^- as the objective function we get U3 as the optimal solution. For the objective function s_{25}^- we get U1 as the solution and last, for the objective function s_5^+ we get also U1 as the solution. Hence our procedure has identified U1, U3 and U4 as alternative optimal solutions, but has not been able to identify the remaining one, U2. This is a very easy example. In practice it is difficult that optimal solutions are single points. In general, solving linear programs as proposed in (2.42) and (2.43) is the sensible way we propose. A final point is worth mentioning. Under VRS, once we have identified three different optimal solutions, all the convex combinations of them are also optimal, which means that we have generated a non-finite number of optimal solutions.

How to Search for Alternative Optimal Solutions When Using a DEA Model with the Same Restrictions as the Additive Model

The last proposed method designed for the weighted additive model can be extended to any other DEA model with the same set of restrictions as the additive model as long as the last added restriction is linear or can be linearized. This happens, for instance, with the "slack-based measure" (Tone, 2001), which is equivalent to the "enhanced Russell graph measure" (Pastor et al. 1999), whose objective function is not linear but fractional. In this case the added non-linear restriction is $\frac{1-\frac{1}{m}\sum_{i=1}^{m}\frac{s_i^-}{x_{i0}}}{1+\frac{1}{s}\sum_{r=1}^{s}\frac{s_r^+}{y_{r0}}} = v^*$, that can be linearized just by transposing its left hand-side denominator. The same happens with the translation invariant measure proposed by Sharp et al. (2007).

References

Aigner DJ, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6:21–37

Aparicio J, Borras F, Pastor JT, Vidal F (2013) Accounting for slacks to measure and decompose revenue efficiency in the Spanish designation of origin wines with DEA. Eur J Oper Res 231:443–451

Aparicio J, Borras F, Pastor JT, Vidal F (2015) Measuring and decomposing firm's revenue and cost efficiency: the Russell measures revisited. Int J Prod Econ 165:19–28

Aparicio J, Pastor JT, Vidal F (2016) The directional distance function and the translation invariance property. Omega 58:1–3

Asmild M, Pastor JT (2010) Slack free MEA and RDM with comprehensive efficiency measures. Omega 38(6):475–483

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30:1078–1092

Battesse GE, Corra GS (1977) Estimation of a production frontier model: with an application to the pastoral zone of Eastern Australia. Austr J Agric Econ 21(3):169–179

Bogetoft P, Hougaard JL (1999) Efficiency evaluations based on potential (non-proportional) improvements. J Prod Anal 12(3):233–247

Briec W, Lesourd JB (1999) Metric distance function and profit: some duality results. J Optim Theory Appl 101(1):15–33

Briec W, Garderes P (2004) Generalized benefit functions and measurement of utility. Math Methods Oper Res 60:101–123

Chambers RG, Chung Y, Färe R (1996) Benefit and distance functions. J Econ Theory 70:407–419

Chambers RG, Chung Y, Färe R (1998) Profit, directional distance functions, and Nerlovian efficiency. J Optim Theory Appl 98(2):351–364

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

Charnes A, Cooper WW (1984) The non-Archimedean CCR ratio for efficiency analysis: a rejoinder to Boyd and Färe. Eur J Oper Res 15:333–334

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2(6):429–444

Charnes A, Cooper WW, Rhodes E (1979) Short communication: measuring efficiency of decision making units. Eur J Oper Res 3:339

Charnes A, Cooper WW, Golany B, Seiford L, Stutz J (1985) Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. J Econometrics 30:91–107

Cobb CW, Douglas PH (1928) A theory of production. Am Econ Rev 18(Supplement):139–165

Cooper WW, Park KS, Pastor JT (1999) RAM: a range adjusted measure of inefficiency for use with additive models, and relations to others models and measures in DEA. J Prod Anal 11:5–42

Cooper WW, Pastor JT, Borras F, Aparicio J, Pastor D (2011a) BAM: a bounded adjusted measure of efficiency for use with bounded additive models. J Prod Anal 35(2):85–94

Cooper WW, Pastor JT, Aparicio J, Borras F (2011b) Decomposing profit inefficiency in DEA through the weighted additive model. Eur J Oper Res 212(2):411–416

Debreu G (1951) The coefficient of resource utilization. Econometrica 19:273–292

Färe R, Grosskopf S, Lovell CAK (1985) The measurement of efficiency of production, Kluwer Nijhof Publishing

Färe R, Primont D (1995) Multi-output production and duality: theory and applications. Kluwer Academic, Boston

Farrell MJ (1957) The measurement of productive efficiency. J R Statist Soc Ser A Gen 120:253–281

Fukuyama H, Weber WL (2009) A directional slacks-based measure of technical inefficiency. Socio-Economic Plann Sci 43(4):274–287
Lovell CAK, Pastor JT (1995) Units invariant and translation invariant DEA models. Oper Res Lett 18:147–151
Luenberger DG (1992a) Benefit functions and duality. J Math Econ 21:461–481
Luenberger DG (1992b) New optimality principles for economic efficiency and equilibrium. J Optim Theory Appl 75:221–264
Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb-Douglas production functions with composed error. Inter Econ Rev 18:435–444
Nerlove, M. (1965) Estimation and identification of Cobb-Douglas production functions. Rand McNally Company
Pastor JT, Ruiz JL, Sirvent I (1999) An enhanced DEA Russell graph efficiency measure. Eur J Oper Res 115:596–607
Pastor JT, Aparicio J (2010) A note on "a directional slacks-based measure of technical inefficiency". Socio-Economic Plann Sci 44:174–175
Pastor JT, Lovell CAK, Aparicio J (2012) Families of linear efficiency programs based on Debreu's loss function. J Prod Anal 38:109–120
Ray SC (2004) Data envelopment analysis. Theory and techniques for economics and operations research. Cambridge University Press, Boston
Sharp JA, Meng W, Liu W (2007) A modified slacks-based measure model for data envelopment analysis with 'natural' negative outputs and inputs. J Oper Res Soc 58:1672–1677
Shephard RW (1953) Cost and production functions. Princeton University Press, Princeton
Silva Portela MCA, Thanassoulis E, Simpson G (2004) Negative aata in DEA: a directional distance approach applied to bank branches. J Oper Res Soc 55:1111–1121
Tone K (2001) A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res 130:498–509
Zofio JL, Pastor JT, Aparicio J (2013) The directional profit efficiency measure: on why profit inefficiency is either technical or allocative. J Prod Anal 40:257–266

Chapter 3
On Measuring Technological Possibilities by Hypervolumes

Mette Asmild and Jens Leth Hougaard

Abstract Measuring technological possibilities is a somewhat neglected topic in the productivity analysis literature. We discuss existing methods as well as an obvious alternative measure based on hypervolumes. We illustrate the use of a volume-based measure on an empirical case of demolition projects from two different companies and suggest ways of overcoming some issues related to the practical implementation. Finally, we discuss pros and cons of the various approaches.

Keywords Program efficiency · Group comparisons · Technology index · Technological possibilities · Hypervolume

3.1 Introduction

In 2008 a major Danish demolition firm acquired another smaller, but presumably more efficient demolition firm. The question is now whether the former firm actually acquired a superior technology by this acquisition or whether the smaller firm was simply more efficient given the same, or possibly even worse, technological possibilities.

This is but one example of a question often relevant to productivity studies, where it might be interesting to determine whether one production technology, represented by a subsample of observed production plans, is superior to another technology (represented by another subsample of feasible production plans). One might, for example, question whether the regulatory framework of US banks provides better production possibilities than that of their European counterparts, whether the railway reforms in Europe have resulted in improved production possibilities over time for railway operations or whether one organizational form is

M. Asmild (✉) · J.L. Hougaard
IFRO, University of Copenhagen, Copenhagen, Denmark
e-mail: meas@ifro.ku.dk

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_3

better than another, for instance, whether cooperatives offer better production possibilities than investor owned firms.

In the literature on productivity and efficiency analysis surprisingly little attention has been devoted to the issue of quantifying (differences in) technological possibilities. In the present paper we discuss and criticize the existing approaches and introduce an obvious alternative candidate based on hypervolumes.

We address practical challenges such as volume calculation of free disposal as well as convex hulls of the data points and the associated problem of sample size bias when comparing the index values for differently sized subsamples. This enables us to answer our opening question of whether one company provides superior technological possibilities to another. Our subsequent empirical analysis of the demolition projects of the two companies reveals that the acquisition did not, in fact, provide a superior technology but instead the projects in the acquired company had a higher average efficiency, likely due to better management.

3.2 Motivation and Relation to the Literature

Within the non-parametric efficiency literature (see e.g., Cooper et al. 2007) production technologies are represented by estimated frontiers and efficiency of observed production plans is measured relative to these.

A seminal paper by Charnes et al. (1981), suggests a method to determine whether one group of observations is superior to another. First, they identify and then remove what is termed "managerial inefficiency" (i.e., within-group technical inefficiency) of each observation. Second, the adjusted observations from the different groups are compared to the production frontier estimated from the pooled set of observations. The distributions of efficiency scores from each group are then compared in order to evaluate so-called "program efficiency" (or, more generally, group efficiency). In this specific case it is concluded that the one program (PFT) "*has not demonstrated its superior efficiency*" (p. 688 op. cit.) despite involving additional expenditures and therefore should not be preferred. Having reached that conclusion, no attempt to define global measures of technological possibilities was made. Charnes et al. do, however, mention the potential use of stochastic dominance as a way to order the group efficiency distributions.

Practical use of a pooled frontier is not problem-free though. If the technology is assumed to be convex as, for example, in Data Envelopment Analysis (DEA), it is important to notice that the use of a pooled frontier requires that it is meaningful to assume convexity not only *within* the groups but also *between* the groups. Clearly this may be questionable if, for instance, units operate under two different regulatory frameworks, and these may often be the very cases that are interesting to compare.

Subsequent to Charnes et al. (1981), only few theoretical developments have considered how to compare the technological possibilities of one group of observations relative to another. Brocket and Golany (1996) develop non-parametric rank

statistics for comparing program efficiency distributions. Cummins et al. (1999) introduce what they call cross-frontier analysis, where one group's observations are compared to the other group's frontier relying on an assumption of constant returns to scale. Note that comparing observations to a frontier defined from a different set of observations necessitates the use of so-called super-efficiency scores, cf. Andersen and Petersen (1993). For further details of these and related approaches to measuring group performance see e.g. the review in Camanho and Dyson (2006).

The approach we develop in the following does *not* depend on assumptions of convexity nor does it rely on the notion of a pooled frontier. However, as we see it, the main problem of the approaches mentioned above can be illustrated by the following simple example:

Example 1 Consider the figure below illustrating a case of two groups of observations (sub-samples) $A = \{a_1, a_2\}$ and $B = \{b_1, b_2\}$ in the 2-input space (for fixed level of output).

First we notice that all observations are managerially efficient relative to the estimated technology for their own group when measured by the radial Farrell efficiency index (Farrell 1957).[1]

If we compare the observations to a pooled frontier (of the data set $A \cup B$) it can be noticed that observation b_2 is more inefficient than observation a_1, whilst b_1 and a_2 are both on the pooled frontier, and consequently group A appears to be superior to group B.

Moreover, the cross-frontier analysis of Cummins et al. (1999) yields the same conclusion since b_1 and a_2 have the same super efficiency scores relative to the other group's frontier whilst b_2 is less efficient relative to A's frontier than a_1 is relative to B's frontier.

Now, say that we move one observation, here $b_2 \in B$, to group A. From Fig. 3.1 it is clear that this does not change the estimated technologies for neither group A nor B since b_2 is (weakly) dominated relative to both groups. In other words, moving observation b_2 to group A should not affect the conclusion that the possibilities in group A are superior to those in group B.

Yet, using the program efficiency approach we now get that group B becomes superior to group A since B now has all its observations (b_1) on the pooled frontier while A contains both a_1 and b_2 which are inefficient relative to the pooled frontier, also if b_2 is first made technically efficient relative to A's frontier.

Similarly, for the cross-frontier analysis we find that B is now superior to A since B's mean advantage over A is larger than A's mean advantage over B as measured by the traditional radial efficiency and super-efficiency indexes.

We submit that moving one observation (which represents a production possibility) from the inferior group to the superior group should not make the superior group worse nor make the inferior group better in terms of technological

[1]The fact that observation b_2 is a dominated boundary point is not of importance here—it could have been strongly efficient as well. The example is chosen such as to make our argument as simple and clear as possible.

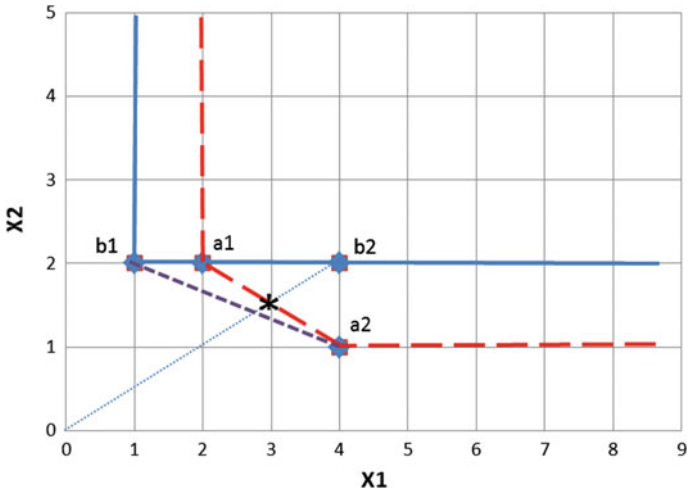


Fig. 3.1 Subsets *A* and *B* in a 2-input space (for the same output level)

possibilities. In other words, adding a possibility can never make a group worse (since either it is dominated by something already observed and adds no new possibilities, or it is undominated and thereby improves what has been observed so far) and likewise removing a possibility can never make a group better in terms of technological possibilities. Δ

Another strand of the literature investigates differences between technologies for subgroups within a dataset by considering variations of the Malmquist index and in particular its frontier shift component (see e.g. Färe et al. 1994, 1998). Here the efficiencies of observed production plans are estimated relative to the frontiers for the different time periods and the ratio of a point's distances to the different frontiers indicates the extent to which the frontier has shifted between the time periods. The standard Malmquist index approach is typically limited to studying changes over time within balanced panel data, but the recently developed Global Malmquist Index, and in particular its global frontier shift component, can be used more generally to compare frontier differences between any subgroups within a data set, see Asmild and Tam (2007).

Note, however, that the use of Malmquist indexes requires an assumption of constant returns to scale of the production technology in order to ensure that the mixed period (or, more generally, cross-frontier) radial efficiency scores of all observations are well defined. Alternatively, one might use a Luenberger indicator based on the traditional directional distance functions, with $g = (x, y)$, which avoids infeasibilities (see Briec and Kerstens 2009) yet this may yield inconsistent results according to Aparicio et al. (2013). Either way, frontier difference measures depend on the choice of efficiency index, and one might argue that comparisons of production *possibilities* should be independent of the measurement of efficiency of individual production plans.

Since we aim to measure the extent (or size) of the production possibilities an index related to the volume of production possibility sets seems an obvious candidate and has, in fact, also been analyzed in the social choice literature related to the notion of economic freedom (e.g., Kolm 2010; Savaglio and Vannucci 2009).

3.3 The Hypervolume Index

Let $x \in \mathbf{R}^s$ be an s-dimensional vector of inputs, let $y \in \mathbf{R}^t$ be a t-dimensional vector of outputs and let $z = (x, y) \in \mathbf{R}^s \times \mathbf{R}^t$ be a feasible production plan, i.e., x can produce y. Let Z be a finite set of such feasible production plans z. Moreover, let

$$\widehat{Z} = (z^{min} + \mathbf{R}_+^{s+t}) \cap (z^{max} - \mathbf{R}_+^{s+t})$$

where

$$z^{min} = (\min_Z\{z_1\}, \ldots, \min_Z\{z_{s+t}\}) \text{ and } z^{max} = (\max_Z\{z_1\}, \ldots, \max_Z\{z_{s+t}\}),$$

be the smallest $s+t$-dimensional hypercube containing Z.

For $A \subseteq Z$ let $D(A) = \{(x, y) \in \mathbf{R}^s \times \mathbf{R}^t | \exists (x_a, y_a) \in A : x_a \leq x, y_a \geq y\}$ be the set of points dominated by some point in A, i.e., the *dominance set* of A.

Likewise, for any subset $A \subseteq Z$ and the hypercube $\widehat{Z}$, we define the *hypercube-restricted free disposal hull* of A as

$$D(A, \widehat{Z}) = D(A) \cap \widehat{Z}. \tag{3.1}$$

Now, for any subset $A \subseteq Z$ and the hypercube $\widehat{Z}$ let $(A, \widehat{Z})$ denote a *problem* (of technological possibilities). In particular we say that a problem is *well-behaved* if $D(A, \widehat{Z}) \subset \mathbf{R}^{s+t}$. Denote by $\mathbb{Z}$ the set of all possible well-behaved problems.

A *technology index* on $\mathbb{Z}$, associates with each well-behaved problem $(A, \widehat{Z})$ an index value $I(A, \widehat{Z}) \in \mathbf{R}_+$.

Let A and B be subsets of a given set Z. If $I(A, \widehat{Z}) \geq (>) I(B, \widehat{Z})$ we say that A offers weakly (strictly) *better production possibilities* than B (given Z).

Given the hypercube $\widehat{Z}$, define a Lebesgue-type volume-based technology index I^V as follows:

$$I^V(A, \widehat{Z}) = \frac{Vol(D(A, \widehat{Z}))}{Vol(\widehat{Z})} \in [0, 1], \tag{3.2}$$

where $Vol(\cdot)$ is the volume operator. Note that $(A, \widehat{Z}) \notin \mathbb{Z} \Rightarrow I^V(A, \widehat{Z}) = 0$, i.e., for problems that are not well-behaved the volume-based index is 0.

Since for arbitrary subsets $A, B \subset Z$,

$$Vol(D(A)) = Vol(D(A)\backslash D(B)) + Vol(D(A) \cap D(B)),$$

we have that,

$$I^V(A, \widehat{Z}) > I^V(B, \widehat{Z}) \Leftrightarrow Vol([D(A)\backslash D(B)] \cap \widehat{Z}) > Vol([D(B)\backslash D(A)] \cap \widehat{Z}. \quad (3.3)$$

Thus, if the volume-based index I^V is higher for A than for B this is equivalent to the volume of A's "advantage" over B being greater than the volume of B's "advantage" over A. Note that the global frontier shift measure of Asmild and Tam (2007) is in fact a discrete (density weighted) approximation of the difference in these "advantages".

Finally, we note that the above framework easily extends to situations where further assumptions on the underlying technologies, for example convexity, are made. As will become clear below, assuming convexity actually simplifies the involved volume calculations.

3.4 Practical Aspects

When it comes to practical application of the index I^V two main issues arise: (1) Computation of volumes and; (2) Sample size bias.

3.4.1 Volume Calculation

Free disposal hull computations While the volumes involved in definition (3.2) are well defined in a mathematical sense, practical calculation of these volumes of box-restricted free disposal hulls in multidimensional spaces is rather cumbersome. Yet, by now there exist several algorithms including the one in Knowles and Corne (2003) with estimated time complexity $O(k^{s+t+1})$ where $k = |A \cap E(A)|$, i.e., the number of undominated elements in A; and the one in Fleischer (2002), with polynomial time complexity $O(k^3(s+t)^2)$.
The algorithm in Fleischer (2002) builds on the natural idea of successively lopping off (parts of) dominance sets and adding these volumes to a partial sum. The procedure then ends when there is no more volume to lop off. While such an algorithm is for exact calculation there also exist faster approximation algorithms based on Monte Carlo sampling, see e.g., Bader and Zitzler (2011).

Convex hull computations In case the volumes are determined by convex envelopment of the data points (instead of free disposal hulls), the volumes can be determined using the software QHULL building on the Quickhull algorithm by

Barber et al. (1996). This does, however, require that all extreme points are provided, which for a variable returns to scale model as in our empirical illustration below, means the (efficient) observations as well as the intersections between their dominance sets and $\widehat{Z}$.

3.4.2 Sample Size Bias

While the hypervolume technology index enable a direct comparison of any two sub-samples, it is worth noting that the index value will be biased in the sense that sets with larger number of observations are more likely to get a larger index value. To investigate the extent of this potential problem we first consider a Monte Carlo simulation study before outlining a practical (resampling based) solution procedure.

Monte Carlo illustration To illustrate the problem of sample size bias we have performed the following Monte Carlo simulation study: Consider the 1-input, 1-output case where the output is uniformly distributed, $U[0, 1]$, the true production technology is given by a Cobb-Douglas function $y = x^2$ and the input inefficiency follows a half-normal distribution. That is, for each observation, first draw an output level and find the corresponding true (efficient) input level. Next, we draw an inefficiency level and add this to the efficient input level.

For the different numbers of observations in the two sub-samples A and B shown in Table 3.1, we performed the above procedure for each observation. For each sub-sample we have computed the index value I^V defined in (3.2) and comparing these shows which sub-sample would appear to offer better production possibilities. Repeating this 10,000 times for each cell in the table we then summarize the proportion of times that sub-sample A is deemed better than sub-sample B.

Since the observations in the two sub-samples are drawn from the same underlying technology we would expect that sub-sample A should be preferred 50%

Table 3.1 Monte Carlo simulation (average from 10000 simulations in each cell) of the percentage of times A is preferred to B for different sample sizes

	# of obs. in B									
		20	30	40	50	60	70	80	90	100
# of obs. in A	20	0.499	0.277	0.150	0.083	0.044	0.022	0.011	0.008	0.005
	30		0.496	0.325	0.208	0.132	0.078	0.050	0.031	0.021
	40			0.498	0.360	0.250	0.162	0.112	0.077	0.055
	50				0.499	0.382	0.282	0.208	0.151	0.109
	60					0.505	0.391	0.307	0.230	0.176
	70						0.501	0.405	0.332	0.254
	80							0.504	0.412	0.343
	90								0.509	0.420
	100									0.501

of the time if there was no sample size bias. Indeed, this can be seen from the diagonal in the table where the sub-samples are of the same size and we see values very close to 0.5. However, moving away from the diagonal we see that the probability of sub-sample A being deemed better than B decreases rapidly when sub-sample A becomes smaller than sub-sample B.

Resampling approach To overcome the problem of sample size bias evident from the simulation results presented above, we suggest the following simple approach based on resampling from the larger sample size: Let n_A and n_B denote the number of observations in subsamples A and B respectively and assume, without loss of generality, that $n_A > n_B$. Then we suggest to select n_B observations from subsample A, calculate the corresponding volume-based index and repeat this a large number of times to arrive at an empirical distribution for the index value for A if this sample had been of the same size as B, in effect doing a "delete $(n_A - n_B)$ jackknife". From the resulting distribution of the sample-size reduced index for A, we can find the average (sample-size reduced) index value, and a corresponding confidence interval determined by the appropriate quantiles in this empirical distribution, which can then be compared to the index value for subsample B.

Alternatively one could consider bootstrapping, with replacement, n_B observations from both A and B and compare their resulting means/distributions. For details of jackknifing and bootstrapping see e.g., Efron (1982) and Shao and Tu (1995).

3.5 Empirical Illustration

To illustrate the suggested approach we consider the empirical case of a large Danish demolition company acquiring a smaller competitor mentioned in the introduction. In order to determine whether this acquisition provided access to a superior production technology, we compare the two subsets of projects undertaken by the two companies. The projects of both companies are described by the same three inputs (machine costs, labour costs and other costs) and the same single output (revenue), for which the averages are shown in Table 3.2.

A straightforward way of comparing the two subsets of projects would be to compare their average efficiencies relative to the pooled frontier, which in a DEA model with variable returns to scale (the BCC model of Banker et al. 1984) reveals that the 34 projects from the smaller newly acquired company has an average

Table 3.2 Numbers of projects and mean costs and revenues for the mother company and the acquired company

	# projects	Machine costs	Labour costs	Other costs	Revenue
Mother	135	293,973	381,696	1,019,005	1,988,614
Acquired	34	33,067	129,145	466,932	741,470
Total	169	241,483	330,887	907,937	1,737,710

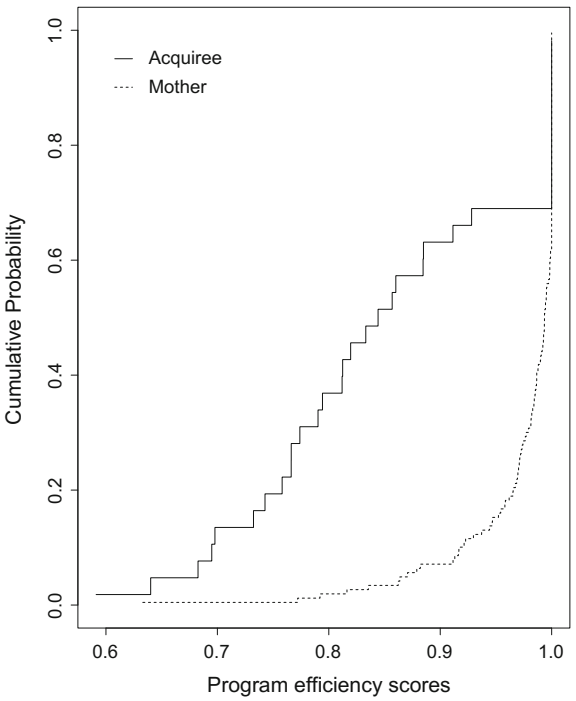


Fig. 3.2 Empirical cumulative density functions

efficiency score of 71.7% compared to an average of 66.3% for the 135 projects from the mother company. Thus one might be tempted to conclude that the acquisition is superior to its mother, but such a comparison does not consider the technological possibilities represented by the two subgroups of projects. In fact, this can be analyzed using the program efficiency approach of Charnes et al. (1981) where we find that the cumulative program efficiency distribution of the mother company consistently lies below that of the acquired company (cf. Fig. 3.2), i.e., the technology of the mother appears to be superior to that of the acquired company.

It is, however, clear from Table 3.2, that the mother company on average has larger, and more machine intensive, projects than the acquired company. Therefore one might question the appropriateness of using a pooled frontier when estimating the projects' efficiencies and assessing the technological possibilities.

Looking first at the volume-based index we find that the index value for the 34 projects from the acquired company is 0.167 whereas the index value for the 135 projects from the mother company is 0.667. Due to the differences in sample sizes, these two values can not be compared directly and we therefore resample by repeatedly drawing 34 observations from the larger subsample. After 100,000 replications this leads to an average (sample-size reduced) index value of 0.461 with an empirical distribution shown in Fig. 3.3.

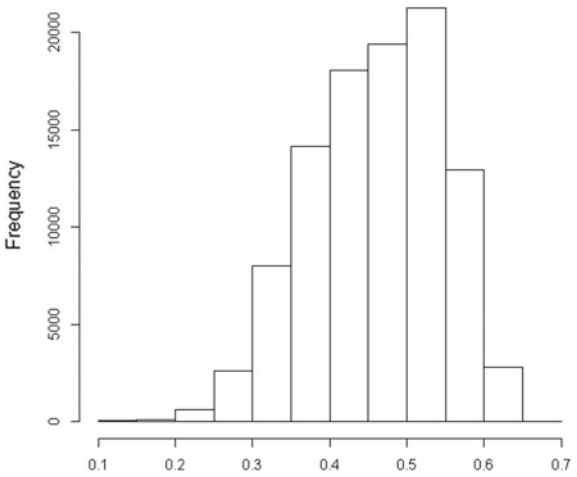


Fig. 3.3 Distribution of sample size reduced hypervolume index values for the mother company

From Fig. 3.3 it is obvious that the average index value for the mother company is larger than that for the acquisition, even if the former is controlled for the sample size bias. This is further confirmed by the empirical 95% confidence interval for the (sample-sized reduced) mean index value of [0.29; 0.60] which does not contain the index value of 0.167 for subsample *B*.

This can then be contrasted to the initial result that the acquisition had a higher average efficiency relative to the pooled frontier. Thus, what was acquired was not a superior technology but rather a better management of the projects in the new company, which the mother company would do well in preserving in the merged company and ideally try to implement on all new projects.

3.6 Discussion

The hypervolume index is in many ways an obvious candidate for measuring the extent of the technological possibilities and above we have illustrated its usage and discussed how to overcome practical aspects of its operationalization.

However, the use of this index is not problem free as the volumes and subsequent ordering of subsets depend on the chosen hypercube $\widehat{Z}$. Here we have chosen to define $\widehat{Z}$ as the smallest hypercube containing all observations but any set covering the observations can in principle be used.

The reason we need the hypercube (or alternative coverings) is first of all so the volumes are bounded and secondly in order to make the index *scale invariant* which clearly is a desirable property. This solution however, implies the dependence on the chosen hypercube. Imagine that an additional observation is added to the data set. Even if this observation is dominated it may still affect the definition of

the hypercube $\widehat{Z}$ and thus the ordering of the subsets. Thus, it violates a property we could call *independence of irrelevant observations*.

Other desirable properties could include a weak version of dominance which can be called *monotonicity in possibilities* meaning that if A has a higher index value than B then moving an observation from B to A cannot result in B getting a higher index value than A. As shown in Example 3.1 in Sect. 3.2 this property is violated by existing methods, yet it is easily shown that this is satisfied by the hypervolume index.

It remains an open question whether it is possible to construct a technology index satisfying even just these desirable properties. So at present the choice of index should be guided by the importance of the various properties for the specific application at hand.

References

Andersen P, Petersen NC (1993) A procedure for ranking efficient units in data envelopment analysis. Manage Sci 39:1261–1294

Aparicio J, Pastor JT, Zofio JL (2013) On the inconsistency of the Malmquist-Luenberger index. Eur J Oper Res 229(3):738–742

Asmild M, Tam F (2007) Estimating global frontier shifts and global Malmquist indices. J Prod Anal 27:137–148

Bader J, Zitzler E (2011) HypE: an algorithm for fast hypervolume-based many-objective optimization. Evol Comput 19:45–76

Banker R, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30:1078–1092

Barber CB, Dobkin DP, Huhdanpaa HT (1996) The quick-hull algorithm for convex hulls. ACM Trans Math Softw 22:469–483

Briec W, Kerstens K (2009) Infeasibility and directional distance functions with application of determinateness of the Luenberger productivity indicator. J Optim Theory Appl 141:55C73

Brocket PL, Golany B (1996) Using rank statistics for determining programmatic efficiency differences in data envelopment analysis. Manage Sci 42:466–472

Camanho AS, Dyson RG (2006) Data envelopment analysis and Malmquist indices for measuring group performance. J Prod Anal 26:35–49

Charnes A, Cooper WW, Rhodes E (1981) Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through. Manage Sci 27:668–697

Cooper WW, Seiford LM, Tone K (2007) Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software, 2nd edn. Springer, Berlin

Cummins JD, Weiss MA, Zi H (1999) Organizational form and efficiency: the coexistence of stock and mutual property-liability insurers. Manage Sci 45:1254–1269

Efron B (1982) The jackknife, the bootstrap and other resampling plans. SIAM, Philadelphia

Färe R, Grosskopf S, Lindgren B, Roos P (1994) Productivity developments in Swedish hospitals: a Malmquist output index approach. In: Charnes A, Cooper WW, Lewin AY, Seiford LM (eds) Data envelopment analysis: theory, methodology and applications. Kluwer Academic Publishers, Dordrecht

Färe R, Grosskopf S, Russell RR (eds) (1998) Index numbers: essays in honor of Sten Malmquist. Kluwer-Academic Publishers, Boston

Farrell MJ (1957) The measurement of productive efficiency. J R Stat Soc A III:253–290

Fleischer M (2002) The measure of Pareto optima: application of multiobjective metaheuristics. CSHCN Technical Report 2002–17

Knowles JD, Corne DW (2003) Properties of an adaptive archiving algorithm for storing nondominated vectors. IEEE Trans Evol Comput 7:100–116

Kolm S-Ch (2010) On real economic freedom. Soc Choice Welfare 35:351–375

Savaglio E, Vannucci S (2009) On the volume-ranking of opportunity sets in economic environments. Soc Choice Welfare 33:1–24

Shao J, Tu D (1995) The jackknife and bootstrap. Springer, Berlin

Chapter 4
Loss Distance Functions and Profit Function: General Duality Results

Juan Aparicio, Fernando Borras, Jesus T. Pastor and Jose L. Zofio

Abstract The concept of loss distance functions is introduced and compared with other functional representations of the technology including the Hölder metric distance functions (Briec and Lesourd in J Optim Theory Appl 101(1):15–33, 1999), the directional distance functions due to Chambers et al. (J Econ Theory 70 (2):407–419 1996; J Optim Theory Appl 98(2):351–364 1998), and the Shephard's input and output distance functions as particular cases of the directional distance functions. Specifically, it is shown that, under appropriate normalization conditions defined over the (intrinsic) input and output prices, the loss distance functions encompass a wide class of both well-known and much less known distance functions. Additionally, a dual correspondence is developed between the loss distance functions and the profit function, and it is shown that all previous dual connections appearing in the literature are special cases of this general correspondence. Finally, we obtain several interesting results assuming differentiability.

Keywords Loss distance functions · Directional distance functions · Hölder distance functions · Duality · Profit function

J. Aparicio (✉) · J.T. Pastor
Center of Operations Research (CIO), Universidad Miguel Hernandez de Elche, 03202, Elche, Alicante, Spain
e-mail: j.aparicio@umh.es

F. Borras
Department of Statistics, Mathematics and Computer Science, University Miguel Hernandez of Elche, Alicante, Spain

J.L. Zofio
Departamento de Analisis Economico: Teoria Economica e Historia Economica, Universidad Autonoma de Madrid, 28049 Madrid, Spain

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_4

4.1 Introduction

The usual starting-point of the modern theory of production is the production possibility set; i.e., all production plans available for the producer. Given a set of postulates about the features of the production possibility set, several functions of great interest from an economic perspective can be derived and characterized. In particular, the analysis related to the theory of the firm leads to a set of well-known functions. We list the most relevant below:

(a) The production function or, in general, a transformation function;
(b) The profit function;
(c) The cost function;
(d) The revenue function;
(e) The output supply function, for each output, and the input demand function, for each input.

The theory of the firm through duality takes any of the above functions as a starting-point and derives the properties of the others as counterparts. In addition, under suitable conditions, it is possible to recover one function from the other one and conversely. So, for example, the duality allows us to establish that the production function contains enough information about the production process so as to derive useful economic information about the firm (profit function, supply and demand functions, etc.). Indeed, duality theory has resulted in two significant breakthroughs in microeconomics. First, due to its econometric implications, duality theory enables us to derive systems of demand and supply functions which are consistent with the behaviour of the producer. Secondly, it allows us to derive the well-known comparative statics theorems as an alternative to those originally deduced by Samuelson (1947).

Based on duality theory, we provide new links between the production model and the optimizing behaviour of the producer. Hotelling (1932) was the first to introduce the profit function and state his famous lemma in the one output case. Hotelling's lemma is the key to obtaining expressions for the input demand and output supply functions assuming simply first order differentiability. Samuelson (1953–54) introduced the concept of the variable profit function and studied some of its properties. Debreu (1959), Gorman (1968) and McFadden (1966, 1978) have also studied the properties of the profit function. In particular, Gorman and McFadden showed that if the production possibility set satisfies certain regularity conditions then the profit function may be used to determine the set of all feasible plans for the firm. In this respect, suitable functional forms for profit functions have been discussed, for example, by Diewert (1973). Diewert (1974, 1982) provides additional references and historical notes.

To represent the production possibility set, it is usual to resort to a simple equation representing the production function, in the single output case, or the transformation function, in the general multi-input multi-output case. The existence of such an equation is not obvious [see Shephard (1970, Chap. 3)]. For this reason,

among others, it seems useful to search for functions that allow identifying substitution alternatives between inputs given an output, or between outputs given an input, in an easy way. These functions are referred as distance functions in the literature.

One of the most important contributions to the distance function representation of the technology is the work pioneered by Shephard (1953, 1970). He defined the well-known input and output distance functions and established several dual relationships. Later, Färe and Primont (1995) developed a dual correspondence between Shephard's distance functions and the profit function. In recent years there has been growing interest in the duality theory and distance functions. First, Luenberger (1992a, b) and later Chambers et al. (1996, 1998) and Briec and Lesourd (1999), have contributed a series of studies along this line. Specifically, Luenberger introduced the concept of the benefit function as a representation of the amount that an individual is willing to trade, in terms of a specific reference commodity bundle, g, for the opportunity to move from a consumption bundle to a utility threshold. Luenberger also defined a so-called shortage function, which basically measures the distance in the direction of a vector g of a production plan from the boundary of the production possibility set. In other words, the shortage function measures the amount by which a specific plan is short of reaching the frontier of the technology. In recent times, Chambers et al. (1996, 1998) redefined the benefit function and the shortage function as efficiency measures, introducing to this end the so-called directional distance function. They showed that the directional distance function encompasses, among others, Shephard's input and output distance functions. They also derived a dual correspondence between the directional distance functions and the profit function that generalized all previous dual relationships. Later, Briec and Lesourd (1999) introduced the so-called Hölder metric distance functions intending to relate the concept of efficiency and the notion of distance in topology. Along this line, they proved that the profit function can be derived from the Hölder metric distance functions and that these distance functions can be recovered from the profit function. Another related recent paper is by Briec and Garderes (2004), who tried to generalize the Luenberger's benefit function. Their generalized benefit function is intimately related to topological norms, constituting a shortfall. In fact, they cannot encompass the case of the benefit function when the reference vector g has some zero components, since then the norm associated with g does not satisfy a basic property: If we denote by $\|\cdot\|_g$ the norm associated with vector g, it does not satisfy $\|z\|_g = 0$, if and only if z is the null vector.

In this study we introduce a new family of distance functions, termed loss distance functions in the spirit of Debreu (1951), and show that they encompass the Hölder metric distance functions, the directional distance functions, and therefore Shephard's input and output distance functions—as they represent particular cases of the directional distance functions, see Chambers et al. (1998). Moreover, our approach allows us to define and to study new distance functions, by simply modifying arbitrarily a set of normalization conditions. Along this line, we will establish a general dual correspondence between the new distance functions and the

profit function, and we will show that *all* previous dual relations proposed in the literature are special cases of this general result.

Most of the ideas about the loss distance functions have been influenced by the seminal 1951 article of Gerard Debreu. Debreu (1951) introduced a well-known radial efficiency measure, which he named the "*coefficient of resource utilization*". He derived this scalar from the much less well-known *dead loss* function that characterizes the monetary value of the inefficiencies, and which is to be minimized. The minimization problem originally proposed by Debreu was $\underset{z,p_z}{\text{Min}}\{p_z(z_0 - z)\}$, where z_0 is a vector representing the actual allocation of resources, z is a vector belonging to the set of optimal allocations and p_z is a vector of the corresponding set of intrinsic or shadow price vectors for z. Debreu named the optimal value of this problem "the magnitude of the loss", and he pointed out that "p_z is affected by an arbitrary positive scalar". The influence of this multiplicative scalar means that the magnitude of the loss can be driven to zero by appropriately scaling all components of p_z. In order to eliminate the arbitrary multiplicative factor affecting all the prices, Debreu proposed to divide the objective function by a price index, reformulating the original problem as

$$\underset{z,p_z}{\text{Min}}\{p_z(z_0 - z)/p_z z\}, \text{ or, equivalently, as } \underset{z,p_z}{\text{Max}}\{p_z z/p_z z_0\}.$$

Additionally, Debreu proved that an optimal solution to the above maximization problem is $z^* = \rho \cdot z_0$, where the scalar $\rho(0 < \rho \leq 1)$ is the coefficient of resource utilization mentioned above.

Debreu studied an economic system consisting of two activities, production and consumption, and allowing for three sources of economic loss: underemployment of resources, inefficiency in production and imperfection of the economic organization. We simplify matters by studying the production activity of an economic system having one source of loss, which Debreu calls "the technical inefficiency of production units." In a production context we can use the minimization of the loss function introduced by Debreu to evaluate the technical efficiency of any producer, assuming that the optimal producers have shadow prices affected by a positive scalar unless a normalization scheme is introduced. Note, however, that dividing the objective function above by a price index, as Debreu did, is not the only way to eliminate the arbitrary multiplicative factor problem. Thus, throughout the paper we will use a set of normalization restrictions on the shadow prices, which will include a wide variety of normalization conditions.

Other more recent references on overall efficiency and duality in a Data Envelopment Analysis context are Cooper et al. (2011), Aparicio and Pastor (2011), Aparicio et al. (2013), Färe et al. (2015), and Aparicio et al. (2015).

The remaining of this paper unfolds as follows. In Sect. 4.2 we lay down the basic assumptions and define the new concept of the loss distance function. In Sect. 4.3, we discuss certain open questions about the normalization set. In Sect. 4.4, we study the basic properties of the loss distance function and establish two main theorems related to duality. Moreover, we show that the Hölder metric

distance functions, the directional distance functions and Shephard's input and output distance functions are particular cases of the loss distance functions, considering a specific normalization set in each case. Finally, we assume differentiability and prove several interesting results in Sect. 4.5. Section 4.6 concludes.

4.2 Loss Distance Functions

An economy consists of a number of agents, the role of each of them being to select a plan of action. In the case of producers, their demand for inputs and supply of outputs. A producer is characterized by the limitations on his/her selection, and by the choice criterion (i.e., economic behaviour). The production plan is constrained to belong to a given production possibility set, which represents essentially his/her limited technological knowledge.

Let $x \in R_+^m$ denote a vector of inputs and $y \in R_+^s$ a vector of outputs; the production possibility set (or technology) T is given by

$$T = \left\{(x, y) \in R_+^{m+s} : x \,\text{can produce}\, y\right\}. \tag{4.2.1}$$

In this paper, we assume that T is a subset of R_+^{m+s} that satisfies the following postulates (see Färe et al. 1985).

(P1) $T \neq \varnothing$;
(P2) $T(x) := \{(u, y) \in T : u \leq x\}$ is bounded $\forall\, x \in R_+^m$;
(P3) $(x, y) \in T$, $(x, -y) \leq (x', -y') \Rightarrow (x', y') \in T$, i.e., inputs and outputs are freely disposable;
(P4) T is a closed set;
(P5) T is a convex set.

The producer choose the production plan in T, for a given set of prices, that maximizes profit, i.e., the sum of all properly discounted anticipated future receipts minus the sum of all properly discounted anticipated future outlays (see Debreu 1959). Hence, we think of the firm as a competitive profit maximizer. In other words, he/she takes prices as fixed and chooses a feasible production plan $(x, y) \in T$ which maximizes his/her profit. The resulting (optimum) profit is a function of the vectors of input and output prices, c and p, denoted here by $\Pi(c, p)$, and formally defined as follows:

Definition 1 Given a vector of input and output prices $(c, p) \in R_+^{m+s}$, and a production possibility set T, then the producer's profit function Π is defined as

$$\Pi(c, p) = \sup_{x,y}\{py - cx : (x, y) \in T\}. \tag{4.2.2}$$

As it is assumed that each producer considers prices as given, we are thinking of firms whose input demands and output supplies are relatively small with respect to

the aggregate demands and supplies and, therefore, do not have market power. In other words, we will work on a perfectly competitive market.

A concept of great interest in microeconomics is that of measuring efficiency. Firms are often interested in knowing whether one can produce more with less. To this respect, measuring technical efficiency is necessary to compare the actual performance of the firm with respect to a certain reference subset of the technology. The concept of the weakly efficient subset of T works as such reference set.

Definition 2 The set $\partial^W(T) = \{(x,y) \in T : (u,-v) < (x,-y) \Rightarrow (u,v) \notin T\}$ is called weakly efficient.

Additionally, there exists a subset of $\partial^W(T)$ which is of interest for economists.

Definition 3 The (strongly) efficient subset of T is defined as

$$\partial^S(T) = \{(x,y) \in T : (u,-v) \leq (x,-y),\ (u,v) \neq (x,y) \Rightarrow (u,v) \notin T\}. \quad (4.2.3)$$

The efficient subset of T is related to the notion of Pareto-efficiency (see Koopmans 1951). Indeed, the efficient subset is made up of all feasible production plans which are not dominated.

To formalize our presentation we assume that the set $\partial^S(T)$ is bounded, representing a suitable technological constraint.[1] It allow us to replace "sup" in (4.2.2) by "max" as we show next. Nevertheless, we first need to prove several lemmas.

Lemma 1 *$T(x)$ is a compact set $\forall x \in R^m_+$.*

Proof $T(x)$, as defined in P2, can be equivalently rewritten as $T \cap \{(u,y) \in R^{m+s} : u \leq x\}$. Then, since both sets are closed (see P4), we have that the intersection is closed as well. Finally, thanks to P2, $T(x)$ is a compact set. ■

Lemma 2 *For any $(x,y) \in T$ there exists $(\tilde{x},\tilde{y}) \in \partial^S(T)$ such that $(\tilde{x},-\tilde{y}) \leq (x,-y)$.*

Proof Let us define the following optimization program:

$$\underset{u,v}{\text{Max}}\{(x-u)1_m + (v-y)1_s : (u,v) \in T(x),\ u \leq x,\ v \geq y\}. \quad (4.2.4)$$

The above maximization program is well defined because the objective function is continuous (linear) and the feasible set is compact, since it is the intersection of the compact set $T(x)$ (Lemma 1) and the closed set $\{(u,v) \in R^{m+s} : u \leq x,\ v \geq y\}$. Therefore, the maximum in (4.2.4) is achieved at some point $(\tilde{x},\tilde{y}) \in T(x) \subset T$ which also satisfies $(\tilde{x},-\tilde{y}) \leq (x,-y)$. Let us suppose that there exists $(\bar{u},\bar{v}) \in T$ such that $(\bar{u},-\bar{v}) \leq (\tilde{x},-\tilde{y})$ and $(\bar{u},\bar{v}) \neq (\tilde{x},\tilde{y})$. It is apparent that $(\bar{u},-\bar{v}) \leq (x,-y)$

[1] Shephard (1970, p. 223), in his classic book "Theory of Cost and Production Functions", used a similar argument to prove that the infimum in the cost function is always achieved on the production possibility set. In other words, "inf" can be changed to "min" in the definition of the cost function.

and $(\bar{u}, \bar{v}) \in T(x)$ since $\bar{u} \leq x$ (see P2). Hence, $(\bar{u}, \bar{v})$ is a feasible point of (4.2.4). Regarding the objective value,

$$(x - \bar{u})1_m + (\bar{v} - y)1_s > (x - \tilde{x})1_m + (\tilde{y} - y)1_s$$

since at least one of the following inequalities hold strictly: $-\bar{u} \geq -\tilde{x}$, $\bar{v} \geq \tilde{y}$. This leads to a contradiction with the fact that $(\tilde{x}, \tilde{y})$ is an optimal solution to (4.2.4). Consequently, $(\tilde{x}, \tilde{y}) \in \partial^S(T)$, which is the point we are seeking for and, therefore, the proof is concluded. ■

Lemma 3 *$T = [cl(\partial^S(T)) + D] \cap R_+^{m+s}$, where $D = \left\{(u, v) \in R_+^m \times \left(-R_+^s\right)\right\}$.*

Proof First of all, we prove that $T = [\partial^S(T) + D] \cap R_+^{m+s}$. Given any $(x, y) \in T$ we know that $\exists(\tilde{x}, \tilde{y}) \in \partial^S(T)$ such that $(\tilde{x}, -\tilde{y}) \leq (x, -y)$ by Lemma 2. Then, defining $u = x - \tilde{x} \geq 0_m$ and $v = y - \tilde{y} \leq 0_s$, we have that $(x, y) = (\tilde{x}, \tilde{y}) + (u, v)$ with $(\tilde{x}, \tilde{y}) \in \partial^S(T)$ and $(u, v) \in D$. Therefore, $T \subset \partial^S(T) + D$. In particular, $T \subset [\partial^S(T) + D] \cap R_+^{m+s}$ since $T \subset R_+^{m+s}$. To prove that $[\partial^S(T) + D] \cap R_+^{m+s} \subset T$, if we have $(\tilde{x}, \tilde{y}) \in \partial^S(T) \subset T$ and $(u, v) \in D$, it is evident that if $(x, y) = (\tilde{x}, \tilde{y}) + (u, v)$ belongs to R_+^{m+s} then $(x, y) \in T$, thanks to the free disposability assumption P3. As a consequence, $[\partial^S(T) + D] \cap R_+^{m+s} \subset T$ and, therefore, $T = [\partial^S(T) + D] \cap R_+^{m+s}$. Finally, we will prove that $T = [cl(\partial^S(T)) + D] \cap R_+^{m+s}$. If $(x, y) \in [cl(\partial^S(T)) + D] \cap R_+^{m+s}$ then $(x, y) = (\tilde{x}, \tilde{y}) + (u, v)$ with $(\tilde{x}, \tilde{y}) \in cl(\partial^S(T))$, $(u, v) \in D$ and $(x, y) \in R_+^{m+s}$. Nevertheless, $(\tilde{x}, \tilde{y}) \in T$ since $cl(\partial^S(T)) \subset T$ thanks to P4. Therefore, $(x, y) \in T$ by P3. Consequently, $[cl(\partial^S(T)) + D] \cap R_+^{m+s} \subset T$. Lastly, note that $[\partial^S(T) + D] \cap R_+^{m+s} \subset [cl(\partial^S(T)) + D] \cap R_+^{m+s}$ and, for this reason, $T \subset [cl(\partial^S(T)) + D] \cap R_+^{m+s}$ since $T = [\partial^S(T) + D] \cap R_+^{m+s}$. This leads to $T = [cl(\partial^S(T)) + D] \cap R_+^{m+s}$. ■

The supremum in (4.2.2) is achievable, thanks to the assumption that the efficient set of T is bounded. It is formally established as follows.

Proposition 1 *Let T be a technology which satisfies P1–P5 and let also assume that the set $\partial^S(T)$ is bounded. Then,*

$$\Pi(c, p) = \max_{x,y}\{py - cx : (x, y) \in T\}, \quad \forall (c, p) \in R_+^{m+s}. \tag{4.2.5}$$

Proof

$$\begin{aligned}\Pi(c, p) &= \sup_{x,y}\{py - cx : (x, y) \in T\} = \sup_{(x,y)\in R_+^{m+s}}\{py - cx : (x, y) \in T\} \\ &= \sup_{(x,y)\in R_+^{m+s}}\left\{py - cx : (x, y) \in \left[cl\left(\partial^S(T)\right) + D\right] \cap R_+^{m+s}\right\} \\ &= \sup_{(x,y)\in R_+^{m+s}}\left\{py - cx : (x, y) \in cl\left(\partial^S(T)\right)\right\}.\end{aligned}$$

Now, since $\partial^S(T)$ is bounded we have that $cl(\partial^S(T))$ is bounded as well. Additionally, $cl(\partial^S(T))$, which is a subset of T thanks to P4, is a compact set. Therefore, the feasible set of the last optimization program is compact. Applying Wierstrass's theorem, the supremum in $\sup_{(x,y)\in R_+^{m+s}} \{py - cx : (x,y) \in cl(\partial^S(T))\}$ and, consequently, in (4.2.2) are achieved on T. ■

Additionally, it is well-known that if $\Pi(c,p)$ is achieved at some point of the set T, then there exists $(x,y) \in \partial^W(T)$ such that $py - cx = \Pi(c,p)$ (see, for example, Varian 1992). In other words, the maximum is really achieved on the weakly efficient subset of T.

On the other hand, postulates P1–P5 are standard and allow to establish a duality between the technology and the profit function (see Färe and Primont 1995; p. 128).

$$T = \left\{(x,y) \in R_+^{m+s} : py - cx \leq \Pi(c,p), \ \ \forall\,(c,p) \in R_+^{m+s}\right\}. \tag{4.2.6}$$

From a mathematical point of view, the above duality relation is really a particularization of a theorem due to Minkowski (1911): every closed convex set can be characterized as the intersection of its supporting halfspaces. In fact, the profit function is known in the mathematical literature as the support function associated with the convex set T (see Rockafellar 1972; p. 112). From an economic point of view, this mathematical theorem helps to establish the duality correspondences between several distance functions and the profit function, among other virtues.

Now, we are ready to introduce the concept of the loss distance function, which is measured with respect to a given normalization set denoted by NS.

Definition 4 Let T be a production possibility set satisfying P1–P5. Let $(x,y) \in R_+^{m+s}$ be an input-output vector and let $NS \subset R_+^{m+s}$. The function $L : R_+^{m+s} \times 2^{R_+^{m+s}} \to [-\infty, +\infty]$ defined by

$$L(x,y;NS) = \inf_{\bar{x},\bar{y},\bar{c},\bar{p}} \left\{(\bar{c},\bar{p})(x-\bar{x},\bar{y}-y) : (\bar{x},\bar{y}) \in \partial^W(T), (\bar{c},\bar{p}) \in Q_{(\bar{x},\bar{y})} \cap NS\right\}, \tag{4.2.7}$$

where $Q_{(\bar{x},\bar{y})} = \left\{(c,p) \in R_+^{m+s} : (c,p) \text{ are shadow prices of} (\bar{x},\bar{y})\right\}$, is the loss distance function.[2]

Alternatively, the loss distance function can be written as

$$L(x,y;NS) = \inf_{\bar{x},\bar{y},\bar{c},\bar{p}} \left\{(\bar{c},\bar{p})(x-\bar{x},\bar{y}-y) : (\bar{c},\bar{p}) \in NS, (\bar{x},\bar{y}) \in R_{(\bar{c},\bar{p})}\right\} \tag{4.2.8}$$

where $R_{(\bar{c},\bar{p})} = \{(x,y) \in T : \bar{p}y - \bar{c}x = \Pi(\bar{c},\bar{p})\}$. This second way of describing the loss distance function will allow us to simplify certain proofs below.

[2]Note the similarities between this definition and the minimization of Debreu's dead loss function.

Depending on the structure of NS and the evaluated production plan (x, y), $L(x, y; NS)$ takes different values. In particular, if $NS = \varnothing$ or if, in general, the optimization programs in (4.2.7) and (4.2.8) are infeasible, then we set $L(x, y; NS) = +\infty$. Note also that the loss distance function presents the same arbitrary multiplicative scalar problem by Debreu (1951). The mission of the set NS considered in (4.2.7) and (4.2.8) is to avoid this problem. We devote the next section to study which regularity properties this type of sets must satisfy. Nevertheless, we remark that NS will be usually defined by means of a finite set of (non-necessarily linear) inequalities or equalities.

Basically, $L(x, y; NS)$ is the distance from (x, y) to the weakly efficient frontier of the technology T, but calculated in terms of the normalization set defined over the shadow price vectors. To obtain a distance with economic meaning we evaluate, as Debreu did, the value of the vector $(x - \bar{x}, \bar{y} - y)$ by the shadow prices $(\bar{c}, \bar{p})$ associated with $(\bar{x}, \bar{y}) \in \partial^W(T)$. Therefore, in economic terms, the loss distance function is the monetary value sacrificed by the firm due to technical inefficiency or, equivalently, it can be seen as a firm's opportunity cost.

4.3 The Normalization Set

Normalizing a vector is a convenient mathematical procedure that normally involves establishing that all feasible vectors have a (certain) norm equal to one. In our stated production framework, the normalization concept should be understood in a weaker way and not necessarily related to a mathematical norm. In general, the set NS should satisfy two basic properties. Firstly, it should provide a representative shadow price vector for the different weakly efficient points, with this vector satisfying the conditions stated below. As an essential feature of the price vectors is that their scale is arbitrary, this property ensures that there exist at least a vector of shadow price values satisfying them along a price ray (see Fig. 4.1). One way to establish a suitable scale for prices is to set a norm equal to one, or multiplying all prices by the appropriate scale factor. However, there exist other alternative ways as we will show later. Secondly, it should avoid the arbitrary multiplicative scalar problem. While the satisfaction of the first property is more an interpretative question than a mathematical issue, for the fulfilment of the second property it is enough to impose the following condition on the normalization set.

(C1) NS is closed and $(0_m, 0_s) \notin NS$.

C1 is a general way to avoid the arbitrary multiplicative scalar problem in (4.2.7) and (4.2.8). Since the null vector does not belong to the closed set NS, the distance from $(0_m, 0_s)$ to NS is strictly positive and we cannot achieve the null vector scaling any $(c, p) \in NS$.

In general, the shadow price vectors are defined over R_+^{m+s}, which is a cone. Hence, if we wish that NS contains at least a "representative" of each ray that belongs to R_+^{m+s} we need to impose an additional regularity condition.

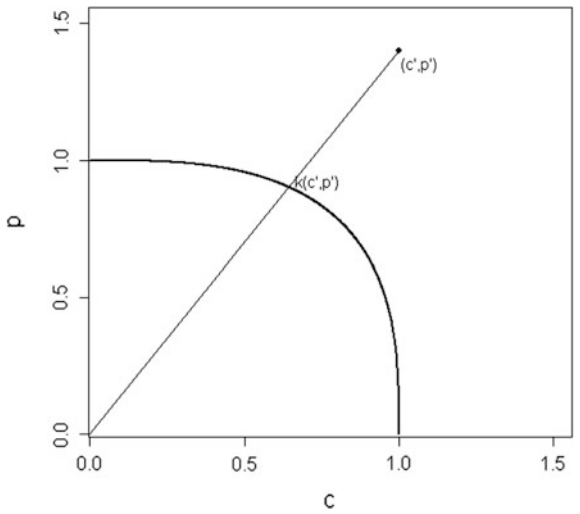


Fig. 4.1 A set satisfying C1 and C2, but does not satisfy the hypothesis of Proposition 2

(C2) $\forall (c,p) \in R_+^{m+s}$, $(c,p) \neq (0_m, 0_s)$, $\exists k > 0$ such that $k \cdot (c,p) \in NS$.

In other words, *NS* satisfying C2 does not rule out any supporting hyperplane of the (closed convex) technology T since each ray is related to this structure. Additionally, we wish to point out that C2 will be very useful for proving several interesting results in the next section. Note also that the constant k above does not have to be unique for a given vector of inputs and outputs prices.

Next we introduce several sufficient requirements for *NS* to prove that such set satisfies C1 and C2. Note that they are sufficient but not necessary conditions.

Proposition 2 *Let NS be a subset of* R_+^{m+s}. *If NS is (1) convex, (2) closed, (3)* $(0_m, 0_s) \notin NS$, *and (4)* $\exists c_i' > 0$ *such that* $\left(0_{i-1}, c_i', 0_{m+s-i}\right) \in NS$, $\forall i = 1, \ldots, m$, *and* $\exists p_r' > 0$ *such that* $\left(0_{m+r-1}, p_r', 0_{m+s-r}\right) \in NS$, $\forall r = 1, \ldots, s$, *then NS satisfies C1 and C2.*

Proof The set *NS* trivially satisfies C1. For C2, it is apparent that $\sum_{i=1}^{m} \lambda_i \left(0_{i-1}, c_i', 0_{m+s-i}\right) + \sum_{r=1}^{s} \mu_r \left(0_{m+r-1}, p_r', 0_{m+s-r}\right)$, with $\lambda_i \geq 0$, $i = 1, \ldots, m$, and $\mu_r \geq 0$, $r = 1, \ldots, s$, generates all R_+^{m+s}. Then, given $(c,p) \in R_+^{m+s}$, $(c,p) \neq (0_m, 0_s)$, $\exists \lambda_i' \geq 0$, $i = 1, \ldots, m$, and $\exists \mu_r' \geq 0$, $r = 1, \ldots, s$, (not all of them zero), such that

$$(c,p) = \sum_{i=1}^{m} \lambda_i' \left(0_{i-1}, c_i', 0_{m+s-i}\right) + \sum_{r=1}^{s} \mu_r' \left(0_{m+r-1}, p_r', 0_{m+s-r}\right).$$

Now, taking $k := \left(\sum_{i=1}^{m} \lambda_i' + \sum_{r=1}^{s} \mu_r'\right)^{-1}$ we get

$$k \cdot (c,p) = \sum_{i=1}^{m} \alpha_i \left(0_{i-1}, c_i', 0_{m+s-i}\right) + \sum_{r=1}^{s} \gamma_r \left(0_{m+r-1}, p_r', 0_{m+s-r}\right)$$

with $\alpha_i = k \cdot \lambda_i' \geq 0$, $i = 1, \ldots, m$, $\gamma_r = k \cdot \mu_r' \geq 0$, $r = 1, \ldots, s$ and $\sum_{i=1}^{m} \alpha_i + \sum_{r=1}^{s} \gamma_r = 1$. Now, by convexity, we have that $k \cdot (c,p) \in NS$. ■

Hence, we have identified a type of sets that satisfy the required conditions.

Example 1 We show several sets complying with the properties in Proposition 2 and, therefore, satisfying C1 and C2:

(a) [A hyperplane] $\left\{(c,p) \in R_+^{m+s} : \alpha_x c + \alpha_y p = \beta\right\}$ with $\alpha = \left(\alpha_x, \alpha_y\right) \in R_{++}^{m+s}$ and $\beta > 0$;

(b) [A polyhedral set] $\left\{(c,p) \in R_+^{m+s} : A \cdot (c,p) \geq b\right\}$, with $A = \left[A_c | A_p\right]$, $A_c = \left[\alpha_x^1 | \ldots | \alpha_x^n\right]$, $\alpha_x^j \in R_{++}^m$, $\forall j = 1, \ldots n$, $A_p = \left[\alpha_y^1 | \ldots | \alpha_y^n\right]$, $\alpha_y^j \in R_{++}^s$, $\forall j = 1, \ldots n$, and $b = (\beta_1, \ldots, \beta_n)$, $\beta_j > 0$, $\forall j = 1, \ldots n$;

(c) [An upper contour set] $\left\{(c,p) \in R_+^{m+s} : f(c,p) \geq \beta\right\}$ with $\beta > 0$ and $f : R_+^{m+s} \to R$ a continuous, quasi-concave function defined such that $f(0_m, 0_s) = 0$, $f\left(0_{i-1}, c_i', 0_{m+s-i}\right) \geq \beta$ for some $c_i' > 0$, $i = 1, \ldots, m$, $f\left(0_{m+r-1}, p_r', 0_{m+s-r}\right) \geq \beta$ for some $p_r' > 0$, $r = 1, \ldots, s$.

Nevertheless, there are sets that do not satisfy the properties of Proposition 2 but satisfies C1 and C2; for example, $NS = \left\{(c,p) \in R_+^{m+s} : \sum_{i=1}^{m} c_i^3 + \sum_{r=1}^{s} p_r^3 = 1\right\}$. Figure 4.1 illustrates it in the one input-one output case.

4.4 Main Results

4.4.1 Characterization of the Technology

We characterize the technology T by means of the loss distance function. We just need to prove two lemmas that require different assumptions.

Lemma 4 *Let T be a subset of R_+^{m+s} that satisfies P1–P5. Let $(x,y) \in R_+^{m+s}$ and let NS be a nonempty subset of R_+^{m+s} that satisfies C1. Then, if $(x,y) \in T$ we have that $L(x,y;NS) \geq 0$.*

Proof If $(x,y) \in T$ then $\bar{p}y - \bar{c}x \leq \Pi(\bar{c},\bar{p})$ $\forall (\bar{c},\bar{p}) \in R_+^{m+s}$ by (4.2.6). Equivalently, $(\bar{p}\,\bar{y} - \bar{c}\,\bar{x}) - (\bar{p}y - \bar{c}x) \geq 0 \quad \forall (\bar{c},\bar{p}) \in R_+^{m+s}$ and $\forall (\bar{x},\bar{y}) \in R_{(\bar{c},\bar{p})}$. Therefore, $(\bar{c},\bar{p})(x - \bar{x}, \bar{y} - y) \geq 0$ $\forall (\bar{c},\bar{p}) \in R_+^{m+s}$ and $\forall (\bar{x},\bar{y}) \in R_{(\bar{c},\bar{p})}$. In particular, it is true $\forall (\bar{c},\bar{p}) \in NS \subset R_+^{m+s}$ and $\forall (\bar{x},\bar{y}) \in R_{(\bar{c},\bar{p})}$. Consequently, the infimum that defines the loss distance function according to (4.2.8) has to be nonnegative. ■

Lemma 5 *Let T be a subset of R_+^{m+s} that satisfies P1–P5. Let $(x,y) \in R_+^{m+s}$ and let NS be a nonempty subset of R_+^{m+s} that satisfies C1 and C2. Then, if $(x,y) \notin T$ we have that $L(x,y;NS) < 0$.*

Proof Suppose that $(x,y) \notin T$. Then, $\exists (c',p') \in R_+^{m+s} \backslash \{(0_m, 0_s)\}$ such that $p'y - c'x > \Pi(c',p')$, by (4.2.6). On the other hand, we know that $\exists (x',y') \in R_{(c',p')}$. Now, by C2, $\exists k > 0$ such that $k \cdot (c',p') \in NS$. It is apparent that $k \cdot (c',p') \in Q_{(x',y')}$, as defined in (4.2.7). Therefore, $k \cdot (c',p')(x - x', y' - y) = k \cdot [p'y' - c'x' - (p'y - c'x)] = k \cdot [\Pi(c',p') - (p'y - c'x)] < 0$ and, finally, $L(x,y;NS) < 0$ according to (4.2.7). ■

Finally, it follows immediately from the above lemmas that it is possible to characterize the production possibility set through the sign of the loss distance function.

Proposition 3 *Let T be a subset of R_+^{m+s} that satisfies P1–P5. Let $(x,y) \in R_+^{m+s}$ and let NS be a nonempty subset of R_+^{m+s} that satisfies C1 and C2. Then, $(x,y) \in T$ if and only if $L(x,y;NS) \geq 0$.*

4.4.2 Properties of the Loss Distance Functions

Now, let us study the properties of the new distance functions introduced in Sect. 4.2.

Proposition 4 *Let T be a subset of R_+^{m+s} that satisfies P1–P5. Let $(x,y), (x',y') \in R_+^{m+s}$ such that $-\infty < L(x,y;NS) < +\infty$ and $-\infty < L(x',y';NS) < +\infty$ and let NS be a nonempty subset of R_+^{m+s} that satisfies C1. Then, the function $L(x,y;NS)$ as defined in (4.2.7) satisfies the following properties.*

(a) $x' \geq x,\ y' \leq y \Rightarrow L(x',y';NS) \geq L(x,y;NS)$;
(b) $L(x,y;NS)$ is concave in (x,y);
(c) $L(x,y;NS)$ is continuous with respect to (x,y) on each open convex subset of R_+^{m+s} in which it is finite;
(d) Let (x,y) be a point where $L(x,y;NS)$ is finite. For each $(\tilde{x},\tilde{y})$, there exists the one-sided directional derivative of $L(x,y;NS)$ at (x,y) with respect to the vector $(\tilde{x},\tilde{y})$.

Proof

(a) Since $x' \geq x \Rightarrow x' - \bar{x} \geq x - \bar{x}$ and $y' \leq y \Rightarrow \bar{y} - y' \geq \bar{y} - y$, we have that

$$(\bar{c},\bar{p})(x' - \bar{x}, \bar{y} - y') \geq (\bar{c},\bar{p})(x - \bar{x}, \bar{y} - y),\ \forall (\bar{x},\bar{y}) \in \partial^W(T), \forall (\bar{c},\bar{p}) \in Q_{(\bar{x},\bar{y})} \cap NS.$$

Therefore, $L(x', y'; NS) \geq L(x, y; NS)$.

(b) Given $(x, y), (x', y') \in R_+^{m+s}$ and $\alpha \in [0, 1]$ we have that
$(\bar{c}, \bar{p})(\alpha x + (1-\alpha)x' - \bar{x}, \bar{y} - \alpha y - (1-\alpha)y') = \alpha((\bar{c}, \bar{p})(x - \bar{x}, \bar{y} - y))$
$+ (1-\alpha)((\bar{c}, \bar{p})(x' - \bar{x}, \bar{y} - y')), \forall (\bar{x}, \bar{y}) \in \partial^W(T), \forall (\bar{c}, \bar{p}) \in Q_{(\bar{x}, \bar{y})} \cap NS.$
Then, since $(\bar{c}, \bar{p})(x - \bar{x}, \bar{y} - y) \geq L(x, y; NS)$ and $(\bar{c}, \bar{p})(x' - \bar{x}, \bar{y} - y') \geq L(x', y'; NS), \forall (\bar{x}, \bar{y}) \in \partial^W(T), \forall (\bar{c}, \bar{p}) \in Q_{(\bar{x}, \bar{y})} \cap NS$, we obtain that

$$
\begin{aligned}
L(\alpha x + (1-\alpha)x', \alpha y + (1-\alpha)y'; NS) &= \inf_{\bar{x}, \bar{y}, \bar{c}, \bar{p}} \{(\bar{c}, \bar{p})(\alpha x + (1-\alpha)x' - \bar{x}, \bar{y} \\
&\quad - \alpha y - (1-\alpha)y') : (\bar{x}, \bar{y}) \in \partial^W(T), (\bar{c}, \bar{p}) \in Q_{(\bar{x}, \bar{y})} \cap NS\} \\
&\geq \alpha L(x, y; NS) + (1-\alpha) L(x', y'; NS).
\end{aligned}
$$

In other words, $L(x, y; NS)$ is concave in (x, y).

(c) The continuity property stated in the proposition is true for any concave function (see Mangasarian 1994; p. 62).

(d) This property is consequence of the concavity of,$L(x, y; NS)$ see Rockafellar (1972; p. 214) for a proof. ■

As a result, the loss distance function satisfies the weak monotonicity condition on R_+^{m+s} and, additionally, it is concave, continuous and one-sided directionally differentiable.

4.4.3 A First Result on Duality

Convexity (P5) is an essential postulate in our work. A convex set has a nonzero normal at each of its boundary points (see Rockafellar 1972; p. 100). Therefore, this property guarantees the existence of intrinsic prices for each boundary point. In fact, it is a standard result that each weakly efficient point of a convex production possibility set has at least a normal with nonnegative coefficients. Additionally, this normal is related to a supporting hyperplane to T and a linear function which achieves its maximum on T. It can be translated in our context as for each $(\bar{x}, \bar{y}) \in \partial^W(T)$ there exists $(\bar{c}, \bar{p}) \in R_+^{m+s}$ such that $\bar{p}\,\bar{y} - \bar{c}\,\bar{x} = \Pi(\bar{c}, \bar{p})$.

Returning now to our initial loss function Definition (4), and since the nonlinear objective function of the corresponding optimization program is equivalent to

$$(\bar{c}, \bar{p})(x - \bar{x}, \bar{y} - y) = (\bar{p}\,\bar{y} - \bar{c}\,\bar{x}) - (\bar{p}y - \bar{c}x) = \Pi(\bar{c}, \bar{p}) - (\bar{p}y - \bar{c}x)$$

for any $(\bar{c}, \bar{p}) \in NS$ and any $(\bar{x}, \bar{y}) \in R_{(\bar{c}, \bar{p})}$, we get the following duality result between the loss distance functions and the profit function.

Theorem 1 *Let T be a subset of R_+^{m+s} that satisfies P1–P5 and let NS be a nonempty subset of R_+^{m+s}. Then,*

$$L(x,y;NS) = \inf_{(c,p)\in R_+^{m+s}} \{\Pi(c,p) - (py - cx) : (c,p) \in NS\}. \tag{4.4.1}$$

As a direct consequence of Theorem 1, we can recover the loss distance functions from the information of the profit function and the set of prices *NS*. Note also that the above result holds independently of whether the set *NS* satisfies C1 and C2, or not.

We can refine the above result in the case in which the production possibility set is a polytope. Throughout the paper we have supposed that the efficient subset of T is bounded. However, if T is a polytope, we do not need to keep boundedness as we show next.

Proposition 5 *Let T be a polytope of R_+^{m+s} that satisfies P1–P5, and let NS be a nonempty subset of R_+^{m+s}. Then,*

$$L(x,y;NS) = \inf_{(c,p)\in R_+^{m+s}} \{\Pi(c,p) - (py - cx) : (c,p) \in NS\}. \tag{4.4.2}$$

Proof If $\forall (c,p) \in NS\ \Pi(c,p) = +\infty$, then the optimization program in (4.2.8) is infeasible and, as a result, $L(x,y;NS) = +\infty$ by definition. Moreover, $\Pi(c,p) - (py - cx) = +\infty$, $\forall (c,p) \in NS$. Therefore, the infimum of $\Pi(c,p) - (py - cx)$ calculated on $(c,p) \in NS$ is equal to $+\infty$. And the proposition holds trivially in this case.

On the other hand, if $\exists (\hat{c},\hat{p}) \in NS$ such that $\Pi(\hat{c},\hat{p}) < +\infty$, and given that if a linear function is bounded from above on a polytope, it achieves its supremum on the polytope (see Mangasarian 1994; p. 130), we have that there exists $(\hat{x},\hat{y}) \in R_{(\hat{c},\hat{p})}$ such that $\Pi(\hat{c},\hat{p}) = \hat{p}\hat{y} - \hat{c}\hat{x}$. Then, it is apparent that $L(x,y;NS) < +\infty$ and $\inf_{(c,p)\in R_+^{m+s}} \{\Pi(c,p) - (py - cx) : (c,p) \in NS\} < +\infty$. Now, let NS^∞ be a set defined as $NS^\infty := \{(c,p) \in NS : \Pi(c,p) = +\infty\}$. Obviously, $NS^\infty \subset NS$ and $NS^\infty \neq NS$. Then, we have that the optimization program in (4.2.8) is equivalent to $\inf_{\bar{x},\bar{y},\bar{c},\bar{p}} \left\{(\bar{c},\bar{p})(x - \bar{x}, \bar{y} - y) : (\bar{c},\bar{p}) \in NS \backslash NS^\infty, (\bar{x},\bar{y}) \in R_{(\bar{c},\bar{p})}\right\}$ and the optimization program in (4.2.8) is equivalent to $\inf_{(c,p)\in R_+^{m+s}} \{\Pi(c,p) - (py - cx) : (c,p) \in NS \backslash NS^\infty\}$. Finally, note that $(\bar{c},\bar{p})(x - \bar{x}, \bar{y} - y) = \Pi(\bar{c},\bar{p}) - (\bar{p}y - \bar{c}x)$, $\forall (\bar{c},\bar{p}) \in NS \backslash NS^\infty$, $\forall (\bar{x},\bar{y}) \in R_{(\bar{c},\bar{p})}$. Therefore, the last two optimization programs are equivalent. ■

Using the method of activity analysis [see Koopmans (1951) and Farrell (1957) or, more recently, Charnes et al. (1978)], a technology can be constructed from n observations of inputs and outputs, interpreting each of them as a feasible production plan. This type of technology defines a feasible region with the shape of a polytope. Following Proposition 5, in these cases the loss distance functions can be recovered from the profit function under a bounded or an unbounded efficient subset.

In order to illustrate the implications of Theorem 1, we can study the following examples where the normalization set will lead to some well-known distance functions.

Example 2 If $NS = \{(c,p) \in R_+^{m+s} : cg_x + pg_y = 1\}$, with $g = (g_x, g_y)$ being a nonzero vector of R_+^{m+s}, given any $(x,y) \in R_+^{m+s}$ we have, by (4.2.8), that

$$\begin{aligned} L(x,y;NS) &= \inf_{(c,p)\in R_+^{m+s}} \{\Pi(c,p) - (py - cx) : cg_x + pg_y = 1\} \\ &= \inf_{(c,p)\in R_+^{m+s}} \left\{ \frac{\Pi(c,p) - (py - cx)}{cg_x + pg_y} \right\} = \vec{D}_T(x,y;g_x,g_y). \end{aligned}$$

Therefore, the loss distance function nests the directional distance function (see Chambers et al. 1998; p. 358) when using a particular normalization set. Also, if we additionally take $g_y = 0_s$ then we obtain the directional input distance functions defined by Chambers et al. (1996), analogue to Luenberger's benefit functions in the production context (see Luenberger 1992a). ■

Therefore, Shephard's input and output distance functions can be also included as a type of loss distance functions, since they are particular cases of the directional distance functions [see Chambers et al. (1998, p. 355)].

Example 3 If $NS = \left\{(c,p) \in R_+^{m+s} : \|(c,p)\|_q \geq 1\right\}$,[3] we have, by (4.2.8), that

$$L(x,y;NS) = \inf_{(c,p)\in R_+^{m+s}} \left\{\Pi(c,p) - (py - cx) : \|(c,p)\|_q \geq 1\right\} = D_T^t(x,y),$$

where $\frac{1}{q} + \frac{1}{t} = 1$, $t = +\infty$ if $q = 1$ and $t = 1$ if $q = +\infty$.

Consequently, the loss distance functions nest Hölder's metric distance functions (see Briec and Lesourd 1999; p. 24), by using a specific normalization condition. ■

We have just shown, by means of Examples 2 and 3, that the loss distance function is an approach that includes a wide class of existing distance functions.

On the other hand, in order to measure efficiency, it is useful to characterize the weakly efficient points. In particular, it would be even more useful if the set of weakly efficient points could be described as the set of zeros of a function. For this, let us observe that the loss distance functions are equal to zero if and only if the evaluated point lies on the weakly efficient frontier of the production possibility set.

Lemma 6 *Let T be a subset of R_+^{m+s} that satisfies P1–P5. Let $(x,y) \in R_+^{m+s}$ and let NS be a nonempty subset of R_+^{m+s} that satisfies C1 and C2. Then, $L(x,y;NS) = 0$ if and only if $(x,y) \in \partial^W(T)$.*

[3] The Hölder norms ℓ_q are defined as $\|(c,p)\|_q = \begin{cases} \left[\sum_{i=1}^m |c_i|^q + \sum_{r=1}^s |p_r|^q\right]^{1/q}, & \text{if } q \in [1,+\infty) \\ \max\{|c_1|,\ldots,|c_m|,|p_1|,\ldots,|p_s|\}, & \text{if } q = +\infty \end{cases}$.

Proof Obviously, the stated result holds for $(x,y) \notin T$ by Proposition 3. Additionally, if $(x,y) \in \partial^W(T)$ there exists $(c,p) \in Q_{(x,y)} \backslash \{(0_m, 0_s)\}$. Hence, by C2, $\exists k > 0$ such that $k \cdot (c,p) \in Q_{(x,y)} \cap NS$. Then, $(\bar{c}, \bar{p})(x - \bar{x}, \bar{y} - y) = 0$ with $(\bar{c}, \bar{p}) = k \cdot (c,p)$ and $(\bar{x}, \bar{y}) = (x,y)$. Finally, by Lemma 4, $L(x,y;NS) = 0$.

Thus, suppose that $(x,y) \in T \backslash \partial^W(T)$. Then, we study two complementary cases:

(i) If $(x,y) \in \text{int}(T)$ there exists $B((x,y),\rho)$, an Euclidean ball of radius ρ centered at (x,y) in R^{m+s}, with $\rho > 0$, such that $B((x,y),\rho) \subset T$. For any $(c,p) \in NS$ we have that the Euclidean distance from the point (x,y) to the supporting hyperplane $p\tilde{y} - c\tilde{x} = \Pi(c,p)$ is equal to $\frac{\Pi(c,p) - (py - cx)}{\sqrt{\sum_{i=1}^m c_i^2 + \sum_{r=1}^s p_r^2}} \geq \rho$. Hence, for any $(c,p) \in NS$ we have that $\Pi(c,p) - (py - cx) \geq \rho \sqrt{\sum_{i=1}^m c_i^2 + \sum_{r=1}^s p_r^2}$. Then, if we calculate the infimum of $\Pi(c,p) - (py - cx)$ on $(c,p) \in NS$, we obtain, thanks to Theorem 1 and the above expression, that

$$L(x,y;NS) \geq \rho \min_{c,p} \left\{ \sqrt{\sum_{i=1}^m c_i^2 + \sum_{r=1}^s p_r^2} : (c,p) \in NS \right\}.$$

The minimization operation above is well defined because the problem calculates the Euclidean distance from $(0_m, 0_s)$ to the closed nonempty set NS. Now, due to C1, we have that $\min_{c,p} \left\{ \sqrt{\sum_{i=1}^m c_i^2 + \sum_{r=1}^s p_r^2} : (c,p) \in NS \right\} > 0$ and, consequently, $L(x,y;NS) > 0$.

(ii) We now suppose that $(x,y) \in T \backslash \partial^W(T)$ but $(x,y) \notin \text{int}(T)$. For this reason, (x,y) has to belong to the boundary of T. However, all points in the boundary of T not generated by the nonnegativity are weakly efficient points thanks to P3 and P4 [see Bonnisseau and Cornet (1988, p. 120) and remember that $T \subset R_+^{m+s}$]. Then, necessarily (x,y) has to belong to the part of the boundary generated by the nonnegativity constraints. If some input of (x,y) were zero, necessarily $(x,y) \in \partial^W(T)$, but it would lead to a contradiction. Therefore, $x_i > 0, i = 1, \ldots, m$, and inevitably $\exists r' = 1, \ldots, s$ such that $y_{r'} = 0$.

On the other hand, since $(x,y) \notin \partial^W(T)$ there exists $(u,v) \in T$ such that $(u,-v) < (x,-y)$, by Definition 2. Thus, $v_r > 0$, $\forall r = 1, \ldots, s$. Now, we can define the point $(\hat{x}, \hat{y}) = \frac{1}{2}(x,y) + \frac{1}{2}(u,v)$, which belongs to T thanks to P5. Then, we can conclude that $(\hat{x}, \hat{y}) \in \text{int}(T)$ since $\hat{y}_r > 0$, $r = 1, \ldots, s$, and at the same time $(u,-v) < (\hat{x}, -\hat{y})$. As a result, we can recall the same arguments in (i) to reach $L(\hat{x}, \hat{y}; NS) > 0$. Finally, by Proposition 4(a), we obtain that $L(x,y;NS) \geq L(\hat{x}, \hat{y}; NS) > 0$ since $(\hat{x}, -\hat{y}) \leq (x, -y)$. ■

Rather than express a production possibility set by means of T, some authors prefer to use the concept of a transformation function in order to describe the set of weakly efficient points. This set can be found described symmetrically (Hicks 1946)

or asymmetrically (Samuelson 1966) in the literature through a transformation function.[4] In our case Lemma 6 constitutes a symmetrical way to describe it as the set of points (x, y) which satisfy the equation $L(x, y; NS) = 0$. Therefore, the loss distance function can be seen as a transformation function.

4.4.4 A Second Duality Result

Shephard (1970) showed that the input distance function is dual to the cost function and the output distance function is dual to the revenue function. Later, Färe and Primont (1995) proved that the profit function and Shephard's distance functions were dual, under several regularity conditions. Chambers, Chung and Färe (1998) established that the profit function can be recovered from the directional distance function, generalizing the dual correspondence between the profit function and Shephard's distance functions. And, finally, Briec and Lesourd (1999) showed that the Hölder metric distance functions are dual to the profit function. In this section, we show that the loss distance functions are (general) precursors of the profit function and indeed both functions are dual. In addition, we prove a general dual correspondence that includes all previous connections appearing in the literature as particular cases.

First of all, we prove a lemma that we will use later in the proof of the main result.

Lemma 7 *Let T be a subset of R_+^{m+s} that satisfies P1–P5. Let us assume that $L(x, y; NS) < +\infty$, $\forall (x, y) \in R_+^{m+s}$. Let NS be a nonempty subset of R_+^{m+s} that satisfies C1. Additionally, let $(c, p) \in NS$. Then,*

$$\Pi(c, p) = \sup_{(x,y)\in R_+^{m+s}} \{py - cx + L(x, y; NS)\}. \tag{4.4.3}$$

Proof By Theorem 1, $\Pi(c, p) - (py - cx) \geq L(x, y; NS)$ for any $(x, y) \in R_+^{m+s}$. Hence, we have that $\Pi(c, p) \geq py - cx + L(x, y; NS)$ for any $(x, y) \in R_+^{m+s}$. Therefore,

$$\Pi(c, p) \geq \sup_{(x,y)\in R_+^{m+s}} \{py - cx + L(x, y; NS)\}.$$

Next we prove the converse. For any $\varepsilon > 0$ $\exists (x, y) \in T$ such that $py - cx \geq \Pi(c, p) - \varepsilon$ (remember that $\Pi(c, p) < +\infty$). Now, by Lemma 4, we have that $L(x, y; NS) \geq 0$ and, therefore, $py - cx + L(x, y; NS) \geq py - cx \geq \Pi(c, p) - \varepsilon$.

[4]The set of weakly efficient points may be described symmetrically as the set of (x, y) satisfying the equation $F(x, y) = 0$, where F is the transformation function. Alternatively, one output can be singled out, for example y_1, and the weakly efficient set may be described asymmetrically by $y_1 = F'(x, y_{-1})$ where F' is the transformation function.

Then, $\sup_{(x,y)\in T}\{py - cx + L(x,y;NS)\} \geq \Pi(c,p)$ since ε is arbitrarily small. Finally,

$$\sup_{(x,y)\in R_+^{m+s}}\{py - cx + L(x,y;NS)\} \geq \Pi(c,p)$$

since $T \subset R_+^{m+s}$. ■

The following theorem shows a more general dual relation between the profit function and the loss distance functions, since the vector of prices does not have to belong directly to the normalization set. It is just necessary that there exists at least "a representative" of the associated ray which belongs to the normalization set.

Theorem 2 *Let T be a subset of R_+^{m+s} that satisfies P1–P5. Let us assume that $L(x,y;NS) < +\infty$, $\forall (x,y) \in R_+^{m+s}$. Let NS be a nonempty subset of R_+^{m+s} that satisfies C1. Additionally, let $(c,p) \in R_+^{m+s}$ such that $\exists k(c,p) > 0$ with $k(c,p)\cdot(c,p) \in NS$. Then,*

$$\Pi(c,p) = \sup_{(x,y)\in R_+^{m+s}}\left\{py - cx + L(x,y;NS)\cdot k(c,p)^{-1}\right\}. \tag{4.4.4}$$

Proof For the sake of simplicity we define $k := k(c,p)$. By hypothesis, $k\cdot(c,p) \in NS$. Therefore, by Lemma 7, we have that $\Pi(kc,kp) = \sup_{(x,y)\in R_+^{m+s}}\{(kp)y - (kc)x + L(x,y;NS)\}$. Now, $\Pi(kc,kp) = k\cdot\Pi(c,p)$ since the profit function is homogeneous of degree +1. Consequently,

$$\begin{aligned}\Pi(c,p) &= k^{-1}\cdot\Pi(kc,kp) = k^{-1}\sup_{(x,y)\in R_+^{m+s}}\{(kp)y - (kc)x + L(x,y;NS)\}\\ &= \sup_{(x,y)\in R_+^{m+s}}\left\{py - cx + L(x,y;NS)\cdot k^{-1}\right\}.\end{aligned}$$

■

As a by-product of Theorem 2, we can state that wide classes of distance functions, derived from the loss distance functions considering different normalization sets, are dual of the profit function. The above theorem shows a general correspondence between distance functions and the profit function.

In addition, note that the duality result given by the theorem above introduces a bit of degeneracy in our problem, since the constant k does not have to be unique for a given vector of inputs and outputs prices. Obviously, the existence of degeneracy will depend on the structure of the normalization set.

As a direct consequence of Theorem 2, we obtain the next corollary.

Corollary 1 *Let T be a subset of R_+^{m+s} that satisfies P1–P5. Let us assume that $L(x,y;NS) < +\infty$, $\forall (x,y) \in R_+^{m+s}$. Let NS be a nonempty subset of R_+^{m+s} that satisfies C1 and C2. Additionally, let $(c,p) \in R_+^{m+s} \backslash \{(0_m, 0_s)\}$. Then,*

$$\Pi(c,p) = \sup_{(x,y)\in R_+^{m+s}} \left\{ py - cx + L(x,y;NS) \cdot k(c,p)^{-1} \right\} \tag{4.4.5}$$

where $k(c,p) > 0$ is such that $k(c,p) \cdot (c,p) \in NS$.

Again, assuming that the technology defines a feasible region with the shape of a polytope, it is unnecessary that the efficient subset is bounded to prove Theorem 2. We state the corresponding result through the following proposition.

Proposition 6 *Let T be a polytope of R_+^{m+s} that satisfies P1–P5. Let us assume that $L(x,y;NS) < +\infty$, $\forall (x,y) \in R_+^{m+s}$. Let NS be a nonempty subset of R_+^{m+s} that satisfies C1. Additionally, let $(c,p) \in R_+^{m+s}$ such that $\exists k(c,p) > 0$ with $k(c,p) \cdot (c,p) \in NS$. Then,*

$$\Pi(c,p) = \sup_{(x,y)\in R_+^{m+s}} \left\{ py - cx + L(x,y;NS) \cdot k(c,p)^{-1} \right\}.$$

Proof Again, we define $k := k(c,p)$. If $\Pi(c,p) < +\infty$ then it is enough to follow the proofs of Lemma 7 and Theorem 2, since if a linear function is bounded from above on a polytope then it achieves its supremum on the polytope. Thus, we will assume that $\Pi(c,p) = +\infty$. Then, there exists a sequence $(x^n, y^n) \in T$ such that $\lim_{n\to\infty} py^n - cx^n = +\infty$. This implies that $\lim_{n\to\infty} [py^n - cx^n + L(x^n, y^n; NS) \cdot k^{-1}] = +\infty$ since $L(x^n, y^n; NS) \geq 0$, by Lemma 4, and $k > 0$. Therefore, $\sup_{(x,y)\in R_+^{m+s}} \{py - cx + L(x,y;NS) \cdot k^{-1}\} = +\infty$, concluding the proof. ■

In order to illustrate the consequences of Theorem 2, we consider the following examples where we show that our result encompasses the dual correspondences stated previously by other authors.

Example 4 If we take $NS = \{(c,p) \in R_+^{m+s} : cg_x + pg_y = 1\}$, with $g = (g_x, g_y)$ being a nonzero vector of R_+^{m+s}, and given any $(c', p') \in R_{++}^{m+s}$, we have that $k \cdot (c', p') \in NS$ with $k = (c'g_x + p'g_y)^{-1}$. Hence, by Theorem 2 and Example 2, we obtain

$$\begin{aligned} \Pi(c', p') &= \sup_{(x,y)\in R_+^{m+s}} \{p'y - c'x + L(x,y;NS) \cdot (c'g_x + p'g_y)\} \\ &= \sup_{(x,y)\in R_+^{m+s}} \{p'y - c'x + \vec{D}_T(x,y;g_x,g_y) \cdot (c'g_x + p'g_y)\}. \end{aligned}$$

Chambers et al. (1998, p. 357) proved the above duality correspondence between the directional distance functions and the profit function for strictly positive prices. This result is a special case of (4.4.4) as we have just shown. ■

Additionally, Chambers et al. (1998) showed that their duality result generalized the relation between the profit function and Shephard's input and output distance functions. Consequently, the loss distance function generalizes this same duality result as well, as illustrated in Example 4.

Example 5 If we take $NS = \left\{(c,p) \in R_+^{m+s} : \|(c,p)\|_q \geq 1\right\}$, with $q \in [1, +\infty)$, given any $(c',p') \in R_+^{m+s} \backslash \{(0_m, 0_s)\}$, we have that $k \cdot (c',p') \in NS$ with $k = \left(\|(c',p')\|_q\right)^{-1}$. As a consequence, by Theorem 2 and Example 3, we now obtain

$$\begin{aligned}\Pi(c',p') &= \sup_{(x,y)\in R_+^{m+s}} \left\{p'y - c'x + L(x,y;NS) \cdot \|(c',p')\|_q\right\} \\ &= \sup_{(x,y)\in R_+^{m+s}} \left\{p'y - c'x + D_T^t(x,y) \cdot \|(c',p')\|_q\right\}.\end{aligned}$$

■

The duality result in Example 5 is similar, but not identical, to the one obtained by Briec and Lesourd (1999) for the Hölder metric distance functions. In the case of Briec and Lesourd's result, to recover the profit function they need to use the Hölder metric distance function and evaluate all (x,y) in T instead of all (x,y) in R_+^{m+s}. Their duality result is as follows

$$\Pi(c',p') = \sup_{x,y}\left\{p'y - c'x - D_T^t(x,y) : (x,y) \in T\right\}. \tag{4.4.6}$$

Our Theorem 2 is more in accordance with the spirit of a standard duality result: to obtain the distance function we just use the information of the profit function, and to recover the profit function we just use the information of the distance function. Nevertheless, we are also able to generalize (4.4.6) by resorting to the loss distance functions thanks to Lemma 6.

Proposition 7 *Let T be a subset of R_+^{m+s} that satisfies P1–P5. Let us assume that $-\infty < L(x,y;NS) < +\infty$, $\forall (x,y) \in R_+^{m+s}$. Let NS be a nonempty subset of R_+^{m+s} that satisfies C1 and C2. Additionally, let $(c,p) \in R_+^{m+s}$. Then,*

$$\Pi(c,p) = \sup_{x,y}\{py - cx - L(x,y;NS) : (x,y) \in T\}. \tag{4.4.7}$$

Proof Let $(x^*,y^*) \in T \backslash \partial^W(T)$. Then, $\exists (\tilde{x},\tilde{y}) \in \partial^S(T) \subset \partial^W(T)$ such that $(\tilde{x},-\tilde{y}) \leq (x^*,-y^*)$ with $(\tilde{x},\tilde{y}) \neq (x^*,y^*)$, by Lemma 2. Now, by Proposition 4(a), we have that

$$py^* - cx^* - L(x^*, y^*; NS) \leq p\tilde{y} - c\tilde{x} - L(\tilde{x}, \tilde{y}; NS).$$

Then,

$$\begin{aligned}&\sup_{x,y}\{py - cx - L(x, y; NS) : (x, y) \in T\}\\&= \sup_{x,y}\{py - cx - L(x, y; NS) : (x, y) \in \partial^W(T)\}.\end{aligned}$$

Finally, since the profit function is achieved by some weakly efficient vector, we have that

$$\begin{aligned}\Pi(c, p) &= \sup_{x,y}\{py - cx : (x, y) \in \partial^W(T)\}\\&= \sup_{x,y}\{py - cx - L(x, y; NS) : (x, y) \in \partial^W(T)\},\end{aligned}$$

by Lemma 6. ■

Briec and Lesourd (1999) state that "*Thus, the above result* [(4.4.6)] *can be considered as a generalization of that of Chambers, Chung and Färe*". This is because the normalization set corresponding to the directional distance functions is a similar case of the normalization set associated to the Hölder distance functions with $t = +\infty$.[5] However, the directional distance functions normalization set is related to a weighted Hölder norm[6] with weights $g = (g_x, g_y) \in R_+^{m+s} \backslash \{(0_m, 0_s)\}$. Briec and Lesourd did not prove a duality result assuming such a type of weighted norms. Moreover, Chambers, Chung and Färe allow zeros in their directional vector g, and in this case the directional distance functions normalization set is not related to the notion of norm in topology.[7] Therefore, strictly speaking, we think that our chapter constitutes really the first formal generalization of the directional distance functions.

Next we propose a last example, far away from the directional distance functions and from the Hölder metric distance functions, as a way of showing the wide framework of the new distance functions.

Example 6 Let $(\omega_x, \omega_y) \in R_{++}^{m+s}$. Then, if $NS = \{(c, p) \in R_+^{m+s} : (c, p) \geq (\omega_x, \omega_y)\}$, *NS* satisfies trivially C1 and given any $(c', p') \in R_{++}^{m+s}$, we have that $k \cdot (c', p')$ belongs to *NS* with

[5] In that case, $NS = \{(c, p) \in R_+^{m+s} : \|(c, p)\|_1 \geq 1\}$ where $\|(c, p)\|_1 = c1_m + p1_s$.

[6] In particular, a weighted norm ℓ_1.

[7] If the vector g has some zero-components then its norm associated does not satisfy the basic property $\|z\| = 0 \Leftrightarrow z = 0_{m+s}$.

$$k = \begin{cases} 1, \text{ if } c' \geq \omega_x, p' \geq \omega_y \\ \left(\min\left\{\omega_{x1}^{-1}c'_1, \ldots, \omega_{xm}^{-1}c'_m, \omega_{y1}^{-1}p'_1, \ldots, \omega_{ys}^{-1}p'_s\right\}\right)^{-1}, \text{ otherwise} \end{cases}.$$

Hence, by Theorem 2,

$$\Pi(c', p') = \begin{cases} \sup\limits_{(x,y)\in R_+^{m+s}} \{p'y - c'x + L(x, y; NS)\}, \text{ if } c'_i \geq \omega_x, p'_r \geq \omega_y \\ \sup\limits_{(x,y)\in R_+^{m+s}} \{p'y - c'x + L(x, y; NS) \cdot \omega\}, \text{ otherwise} \end{cases}$$

where $\omega := \min\left\{\omega_{x1}^{-1}c'_1, \ldots, \omega_{xm}^{-1}c'_m, \omega_{y1}^{-1}p'_1, \ldots, \omega_{ys}^{-1}p'_s\right\}$. ■

Finally, we would like to summarize the existing dual relation between the profit function and the loss distance functions, which establishes a general duality result.

$$\begin{aligned} L(x, y; NS) &= \inf_{(c,p)\in R_+^{m+s}} \{\Pi(c, p) - (py - cx) : (c, p) \in NS\}, \\ \Pi(c, p) &= \sup_{(x,y)\in R_+^{m+s}} \left\{py - cx + L(x, y; NS) \cdot k(c, p)^{-1}\right\}. \end{aligned} \tag{4.4.8}$$

The first expression says that the loss distance function can be derived from the profit function by minimizing the difference between the profit function and the profit at point (x, y) over all feasible prices satisfying the normalization conditions of *NS*. The second relation establishes that if we start with a loss distance function, we can recover the profit function using an optimization program and a strictly positive constant k, directly related to the price vector (c, p) and *NS*.

4.5 Assuming Differentiability

For the proofs in Sect. 4.4 we did not need to assume differentiability of the loss distance functions. However, assuming differentiability of both the loss distance functions and the profit function allows us to achieve several interesting additional results.

First, we consider the optimization program presented in the first part of (4.4.8), i.e.,

$$L(x, y; NS) = -\sup_{(c,p)\in R_+^{m+s}} \{-\Pi(c, p) + (py - cx) : (c, p) \in NS\}. \tag{4.5.1}$$

To obtain the first order conditions of the above optimization program we need to assume some specific structure on the normalization set *NS*. In fact, in order to work with a standard optimization program we need to assume that *NS* is defined by means of a set of constraints. Under the hypothesis of Proposition 2, the

normalization set is a closed convex set and, as a consequence, it can be characterized as the intersection of its supporting hyperplanes. In other words, we could work with a finite or infinite number of linear constraints and derive a standard finite or semi-infinite optimization program, respectively.

In general, we can assume that the normalization set is defined by a set of linear or nonlinear constraints and develop the analysis that we show next. Nevertheless, for the sake of simplicity, we will restrict our attention to a particular type of a single normalization condition:

$$NS = \left\{(c,p) \in R_+^{m+s} : h(c,p) \geq 1\right\}, \tag{4.5.2}$$

where h is a continuous concave function, homogeneous of degree +1 and such that $h(0_m, 0_s) < 1$. This type of normalization condition includes several interesting examples as we show later.

Here, it should be remarked that NS is a closed set since h is a continuous function. Therefore, the normalization set as defined in (4.5.2) satisfies C1 and avoids the arbitrary multiplicative scalar problem pointed out by Debreu (1951).

Next we want to show that, if it is desired to solve the optimization program that appears in (4.5.1) with NS as in (4.5.2), the optimal value of the Lagrangian multiplier associated with the constraint $h(c,p) \geq 1$ is equal to the loss distance function value, i.e., $L(x,y;NS)$.

Proposition 8 *Let $\Pi(c,p)$ be a differentiable function. Also, let $h : R_+^{m+s} \to R$ be a continuous, concave, homogeneous of degree +1 and differentiable function such that $h(0_m, 0_s) < 1$. Additionally, let θ^* be the optimal value of the Lagrangian multiplier associated with the constraint $h(c,p) \geq 1$ in the optimization program (4.5.1) where NS is as in (4.5.2). If the supremum in (4.5.1) is achieved at some point in R_{++}^{m+s} (interior solutions) and any constraint qualification holds,[8] then $\theta^* = L(x,y;NS)$.*

Proof Under the hypothesis, we want really to solve the following optimization program.

$$\max_{(c,p) \in R_{++}^{m+s}} \{-\Pi(c,p) + (py - cx) : h(c,p) - 1 \geq 0\}. \tag{4.5.3}$$

We consider the Lagrangian for the above problem.

$$\Lambda(c,p,\theta) = -\Pi(c,p) + (py - cx) + \theta \cdot (h(c,p) - 1), \tag{4.5.4}$$

where $c \in R^m$, $p \in R^s$ and $\theta \in R$.

Note that the objective function in (4.5.3) is concave since the profit function is convex in positive prices (see Färe and Primont 1995; p. 125) and $py - cx$ is a

[8]For example, we could assume that Slater's constraint qualification holds, i.e., there exists at least a point $(\tilde{c}, \tilde{p}) \in R_{++}^{m+s}$ such that $h(\tilde{c}, \tilde{p}) > 1$ (see Mangasarian 1994; p. 78).

linear function. Additionally, let us remember that h is a concave function and, as a consequence, $h(c,p)-1$ is a concave function as well.

Then, assuming any constraint qualification, (c^*,p^*) is an optimal solution of (4.5.3) if and only if there exist θ^* such that the following Kuhn-Tucker conditions hold.

$$\begin{aligned}
&(a_i) \quad \frac{\partial \Pi(c^*,p^*)}{\partial c_i} = -x_i + \theta^* \frac{\partial h(c^*,p^*)}{\partial c_i}, \quad i=1,\ldots,m;\\
&(b_r) \quad \frac{\partial \Pi(c^*,p^*)}{\partial p_r} = y_r + \theta^* \frac{\partial h(c^*,p^*)}{\partial p_r}, \quad r=1,\ldots,s;\\
&(c) \quad \theta^* \geq 0;\\
&(d) \quad \theta^* \cdot (h(c^*,p^*)-1) = 0;\\
&(e) \quad (c^*,p^*) \in R^{m+s}_{++}, h(c^*,p^*) \geq 1.
\end{aligned} \tag{4.5.5}$$

Now, we multiple each condition (a_i) by c_i^* and each condition (b_r) by p_r^*. Then, we sum all these conditions and obtain, thanks to the homogeneity of degree +1 of the functions $\Pi(c,p)$ and $h(c,p)$, $\Pi(c^*,p^*) = p^*y - c^*x + \theta^* \cdot h(c^*,p^*)$. Now, by (d), $\Pi(c^*,p^*) = p^*y - c^*x + \theta^*$. Then, following (4.5.1), we have that $L(x,y;NS) = \theta^*$. ■

The set NS as defined in (4.5.2) encompasses several interesting cases. For example, let $h(c,p) = a \cdot \left(\prod_{i=1}^m c_i^{\alpha_i}\right)\left(\prod_{r=1}^s p_r^{\beta_r}\right)$ be a Cobb-Douglas function with $\sum_{i=1}^m \alpha_i + \sum_{r=1}^s \beta_r = 1$. Then, the function h is continuous, concave, homogeneous of degree +1 [see Madden (1986, p. 167)] and, obviously, $h(0,0) = 0 < 1$. Therefore, the set NS includes the Cobb-Douglas function as a particular case. The same can be said with respect the C.E.S. (Constant Elasticity of Substitution) function. If $h(c,p) = a \cdot \left[\sum_{i=1}^m \alpha_i c_i^{-\sigma} + \sum_{r=1}^s \beta_r p_r^{-\sigma}\right]^{-1/\sigma}$ then h is a continuous, concave, homogeneous of degree +1 (see Madden 1986; pp. 171–172) and $h(0,0) = 0 < 1$. Moreover, note that we can develop a result similar to Proposition 8 using $h(c,p) = 1$ instead of $h(c,p) \geq 1$ in the definition of NS. So Proposition 9 includes the case of the directional distance functions normalization condition since it is related to a linear function, which is continuous, concave, homogeneous of degree +1 and differentiable.

4.6 Conclusions

In this paper, we recover the seminal idea of Debreu (1951) of evaluating the "dead loss" in the context of the production theory for measuring inefficiency. Following this idea, we have defined the family of loss distance functions and shown that it is a model that encompasses a wide class of existing distance functions.[9] As a

[9]The same idea was already used by Pastor et al. (2012) in a quite different context, i.e., for modelling the DEA efficiency models in a unified way.

consequence, the Hölder metric distance functions, the directional distance functions and Shephard's distance functions can be seen as a type of firms' opportunity costs, using for the evaluation different normalization conditions defined on the intrinsic prices. Additionally, we proved several interesting properties of the loss distance function. As a main result, we derived a general duality relation between the profit function and the loss distance functions that generalize the ones by Chambers et al. (1996, 1998) and Briec and Lesourd (1999). From a practical perspective, the conditions associated to the normalization set can be considered as a benchmark (or check list) when defining new efficiency measures. These conditions entail a rationale when assessing the suitability of these new measures in terms of duality, as it ensures that a consistent decomposition of economic efficiency considering technical and allocative criteria is theoretically grounded. Finally, assuming differentiability, we were able to derive the well-known Hotelling's Lemma, and we proved that the optimal value of the Lagrangian multiplier associated with a continuous, concave, and homogeneous of degree +1 normalization condition, is equal to the loss distance function value.

References

Aparicio J, Pastor JT (2011) A General input distance function based on opportunity costs. Adv Decis Sci 2011:11 (Article ID 505241)

Aparicio J, Borras F, Pastor JT, Vidal F (2013) Accounting for slacks to measure and decompose revenue efficiency in the Spanish designation of origin wines with DEA. Eur J Oper Res 231:443–451

Aparicio J, Borras F, Pastor JT, Vidal F (2015) Measuring and decomposing firm's revenue and cost efficiency: the Russell measures revisited. Int J Prod Econ 165:19–28

Bonnisseau JM, Cornet B (1988) Existence of equilibria when firms follows bounded losses pricing rules. J Math Econ 17:119–147

Briec W, Garderes P (2004) Generalized benefit functions and measurement of utility. Math Methods Oper Res 60:101–123

Briec W, Lesourd JB (1999) Metric distance function and profit: some duality results. J Optim Theory Appl 101(1):15–33

Chambers RG, Chung Y, Färe R (1996) Benefit and distance functions. J Econ Theory 70(2):407–419

Chambers RG, Chung Y, Färe R (1998) Profit, directional distance functions, and nerlovian efficiency. J Optim Theory Appl 98(2):351–364

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444

Cooper WW, Pastor JT, Aparicio J, Borras F (2011) Decomposing profit inefficiency in DEA through the weighted additive model. Eur J Oper Res 212(2):411–416

Debreu G (1951) The coefficient of resource utilization. Econometrica 19(3):273–292

Debreu G (1959) Theory of value. An axiomatic analysis of economic equilibrium. A cowles foundation monograph. Yale University Press, New Haven

Diewert WE (1973) Functional forms for profit and transformation functions. J Econ Theory 6 (3):284–316

Diewert WE (1974) Applications of duality theory. In: Intrilligator MD, Kendrick DA (eds) Frontiers of quantitative economics, vol II. North Holland, Amsterdam, pp 176–199

Diewert WE (1982) Duality approaches to microeconomic theory. In: Arrow KJ, Intriligator MD (eds) Handbook of mathematical economics, vol II. North Holland, Amsterdam, pp 535–600
Färe R, Primont D (1995) Multi-output production and duality: theory and applications. Kluwer Academic Publishers, Boston
Färe R, Grosskopf S, Lovell CAK (1985) The measurement of efficiency of production. Kluwer Nijhof Publishers, Boston
Färe R, Fukuyama H, Grosskopf S, Zelenyuk V (2015) Decomposing profit efficiency using a slack-based directional distance function. Eur J Oper Res 247(1):335–337
Farrell MJ (1957) The measurement of productive efficiency. J Roy Stat Soc A 120:253–290
Gorman WM (1968) Measuring the quantities of fixed factors. In: Wolfe JN (ed) Value, capital and growth: papers in honour of Sir John Hicks. Aldine Publishing Co., Chicago, pp 141–172
Hicks JR (1946) Value and capital. Clarendon Press, Oxford
Hotelling H (1932) Edgeworth taxation paradox and the nature of supply and demand functions. J Polit Econ 30:552–616
Koopmans TC (1951) Analysis of production as an efficient combination of activities. In: Koopmans TC (ed) Activity analysis of production and allocation. Wiley, New York, pp 33–97
Luenberger DG (1992a) Benefit functions and duality. J Math Econ 21:461–481
Luenberger DG (1992b) New optimality principles for economic efficiency and equilibrium. J Optim Theory Appl 75:221–264
Madden P (1986) Concavity and optimization in microeconomics. Blackwell Publisher, Oxford
Mangasarian OL (1994) Nonlinear programming. Classics in Applied Mathematics, vol 10 (SIAM), Philadelphia
McFadden D (1966) Cost, revenue and profit functions: a cursory review. IBER Working Paper 86, University of California at Berkeley
McFadden D (1978) Cost, revenue, and profit functions. In: Fuss M, McFadden D (eds) Production economics: a dual approach to theory and applications, vol 2. Applications to the theory of production, North Holland, Amsterdam
Minkowski H (1911) Theorie der Konvexen Körper, Gesammelte Abhandlungen II. B.G. Teubner, Leipzig, Berlin
Pastor JT, Lovell CAK, Aparicio J (2012) Families of linear efficiency programs based on Debreu's loss function. J Prod Anal 38:109-120
Rockafellar RT (1972) Convex analysis. Princeton University Press. Second Printing, Princeton
Samuelson PA (1947) Foundations of economic analysis. Harvard University Press, Cambridge
Samuelson PA (1953–54) Prices of factors and goods in general equilibrium. Rev Econ Stud 21: 1–20
Samuelson PA (1966) The fundamental singularity theorem for non-joint production. Int Econ Rev 7:34–41
Shephard RW (1953) Cost and production functions. Princeton University Press, Princeton
Shephard RW (1970) Theory of cost and production functions. Princeton University Press, Princeton
Varian HR (1992) Microeconomic analysis, 3rd edn. W.W. Norton and Company, New York

Chapter 5
Exact Relationships Between Fisher Indexes and Theoretical Indexes

Emili Grifell-Tatjé and C. A. Knox Lovell

Abstract We develop exact relationships between empirical Fisher indexes and their theoretical Malmquist and Konüs counterparts. We begin by using implicit Malmquist price and price recovery indexes to establish exact relationships between empirical Fisher quantity and productivity indexes and theoretical Malmquist quantity and productivity indexes. We then show that Malmquist quantity and productivity indexes and Fisher price and price recovery indexes "almost" satisfy the product test with the relevant value change, and we derive a quantity mix function that ensures satisfaction of the product test. We next use implicit Konüs quantity and productivity indexes to establish exact relationships between empirical Fisher price and price recovery indexes and theoretical Konüs price and price recovery indexes. We then show that Konüs price and price recovery indexes and Fisher quantity and productivity indexes "almost" satisfy the product test with the relevant value change, we derive a price mix function that ensures satisfaction of the product test, and we show that this price mix function differs fundamentally from the quantity mix function relating Malmquist and Fisher indexes.

Keywords Implicit Malmquist and implicit Konüs indexes · Fisher indexes · Quantity mix and price mix functions

JEL Classification Codes C43 · D24 · D61

E. Grifell-Tatjé
Department of Business, Universitat Autònoma de Barcelona, Barcelona, Spain
e-mail: emili.grifell@uab.cat

C.A. Knox Lovell (✉)
CEPA, School of Economics, University of Queensland, Brisbane, Australia
e-mail: k.lovell@uq.edu.au

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_5

5.1 Introduction

Empirical Fisher quantity and productivity indexes differ from theoretical Malmquist quantity and productivity indexes. This matters because Fisher quantity and productivity indexes can be calculated from empirical quantity and price data, and Malmquist quantity and productivity indexes have nice properties. It is important to emphasize at the outset that the Malmquist productivity index we analyze is the ratio of a Malmquist output quantity index to a Malmquist input quantity index. It was suggested by Diewert (1992), who attributed it to Hicks and to Moorsteen, and endorsed by Bjurek (1996), who called it the Malmquist total factor productivity index, and evaluated by O'Donnell (2012), who characterized it as being multiplicatively complete, and therefore decomposable into the product of drivers of productivity change, provided that the output quantity index and the input quantity index are non-negative, non-decreasing scalar-valued functions homogeneous of degree +1. This Malmquist productivity index differs from the more popular index bearing the same name introduced by Caves et al. (1982), which cannot be expressed as the ratio of an output quantity index to an input quantity index, and so is not multiplicatively complete.

Our first objective is to relate empirical Fisher quantity and productivity indexes to theoretical Malmquist quantity and productivity indexes, and to derive economically meaningful functions that characterize the disparities between the two. These functions also characterize the extent to which, and the economic explanation for why, Malmquist quantity and productivity indexes and Fisher price and price recovery[1] indexes fail to satisfy the product test[2] with the relevant value change. The key ingredients in this analysis are *implicit* Malmquist price and price recovery indexes.[3]

Similarly, empirical Fisher price and price recovery indexes differ from theoretical Konus price and price recovery indexes. This also matters, because Fisher price and price recovery indexes also can be calculated from empirical price and quantity data, and Konüs price and price recovery indexes also have nice properties. O'Donnell's multiplicatively complete characterization of the Malmquist productivity index also applies to the Konüs price recovery index, enabling one to

[1]A price recovery index is analogous to a productivity index, but is defined in price space rather than quantity space as the ratio of an output price (instead of quantity) index to an input price (instead of quantity) index. It reflects the ability of producers to "recover" financially from input price increases by raising output prices, or to "recover" financially from output price declines by reducing input prices. A popular macroeconomic example is the terms of trade index, the ratio of a country's export price index to its import price index, which, together with its rate of productivity growth, influences a country's economic welfare.

[2]The product test is a test of whether the product of a quantity index and a price index equals the relevant value change. The test is demanding, and not all quantity and price index pairs satisfy it, e.g., a Törnqvist quantity index and a Törnqvist price index.

[3]An implicit price (quantity) index is the ratio of value change to the corresponding quantity (price) index, and trivially satisfies the product test with the corresponding quantity (price) index.

exhaustively decompose it into the product of drivers of price recovery change, although it remains to be seen whether the decomposition makes economic sense. Nonetheless our second objective is to relate empirical Fisher price and price recovery indexes to theoretical Konüs price and price recovery indexes, and to derive (fundamentally different) economically meaningful functions that characterize the disparities between the two. These functions also characterize the extent to which, and the economic explanation for why, Konüs price and price recovery indexes and Fisher quantity and productivity indexes fail to satisfy the product test with the relevant value change. The key ingredients in this analysis are *implicit* Konüs quantity and productivity indexes.

The literature relating empirical and theoretical index numbers has taken two approaches. One approach seeks restrictions on the structure of production technology, in conjunction with an assumption of optimizing behavior, that *equate* an empirical index with a corresponding theoretical index. Diewert (1992) follows this approach to provide "a strong economic justification" for the use of Fisher quantity and productivity indexes. A second approach imposes relatively weak regularity conditions on production technology, sufficient for duality to hold, augmented with a less restrictive form of optimizing behavior, to establish *approximate* relationships between empirical and theoretical indexes. Balk (1998) makes extensive use of this approach.

Our analysis fits into neither category. It begins with implicit theoretical price and quantity indexes. We use these implicit indexes not as ends themselves, but as means to more important ends, the derivation of functions that link empirical Fisher indexes with theoretical Malmquist and Konüs indexes, and that ensure satisfaction of the analogous product tests. We provide economic intuition behind the content of these functions, which characterize variation in the mix of choice variables, either quantities or prices. Our analysis extends results in Grifell-Tatjé and Lovell (2015).

Our analysis proceeds as follows. In Sect. 5.2 we provide some background to motivate our analysis relating empirical and theoretical index numbers. In Sect. 5.3 we use implicit Malmquist price and price recovery indexes to relate empirical Fisher quantity and productivity indexes to theoretical Malmquist quantity and productivity indexes. We also show that Malmquist quantity and productivity indexes and Fisher price and price recovery indexes "almost" satisfy the product test with the relevant value change, and we derive and provide economic interpretations of quantity mix functions that ensure satisfaction of the product test. In Sect. 5.4 we use implicit Konüs quantity and productivity indexes to relate empirical Fisher price and price recovery indexes to theoretical Konüs price and price recovery indexes. We also show that Konüs price and price recovery indexes and Fisher quantity and productivity indexes "almost" satisfy the product test with the relevant value change, we derive price mix functions that ensure satisfaction of the product test, and we show that these price mix functions differ fundamentally from the analogous quantity mix functions relating Malmquist and Fisher quantity and productivity indexes. Section 5.5 concludes.

5.2 Background

Let $y^t \in R^M_+$ and $x^t \in R^N_+$ be output and input quantity vectors with corresponding price vectors $p^t \in R^M_{++}$ and $w^t \in R^N_{++}$, and let revenue $R^t = p^{tT}y^t$, cost $C^t = w^{tT}x^t$, and profitability (or cost recovery) $\Pi^t = R^t/C^t$, all for two time periods, a base period $t = 0$ and a comparison period $t = 1$. The technology is given by $T^t = \{(y, x)$: x can produce y in period t$\}$, its convex output sets are given by $P^t(x) = \{y: (y, x) \in T^t\}$ with frontiers $IP^t(x) = \{y: y \in P^t(x), \lambda y \notin P^t(x), \lambda > 1\}$, and its convex input sets are given by $L^t(y) = \{x: (x, y) \in T^t\}$ with frontiers $IL^t(y) = \{x: x \in L^t(y), \lambda x \notin L^t(y), \lambda < 1\}$. Output distance functions are defined on the output sets by $D^t_o(x, y) = \min\{\phi > 0: y/\phi \in P^t(x)\} \leq 1 \ \forall \ y \in P^t(x)$, and input distance functions are defined on the input sets by $D^t_i(y, x) = \max\{\theta > 0: x/\theta \in L^t(y)\} \geq 1 \ \forall \ x \in L^t(y)$. Finally revenue frontiers are defined on the output sets by $r^t(x, p) = \max_y\{p^Ty: y \in P^t(x)\} \geq R^t$, and cost frontiers are defined on the input sets by $c^t(y, w) = \min_x\{w^Tx: x \in L^t(y)\} \leq C^t$.

We know from Balk (1998) that our best empirical and theoretical quantity and productivity indexes are related by

$$\begin{aligned} Y_F &\cong Y_M\left(x^1, x^0, y^1, y^0\right) \\ X_F &\cong X_M\left(y^1, y^0, x^1, x^0\right) \\ \frac{Y_F}{X_F} &\cong \frac{Y_M(x^1, x^0, y^1, y^0)}{X_M(y^1, y^0, x^1, x^0)}, \end{aligned} \tag{5.1}$$

where Y_F, X_F and Y_F/X_F are Fisher output quantity, input quantity and productivity indexes, and $Y_M(x^1, x^0, y^1, y^0)$, $X_M(y^1, y^0, x^1, x^0)$ and $Y_M(x^1, x^0, y^1, y^0)/X_M(y^1, y^0, x^1, x^0)$ are Malmquist output quantity, input quantity and productivity indexes in geometric mean form.[4]

It follows that[5]

---

[4]The Fisher quantity indexes are defined as $Y_F = \left[\left(\frac{p^{0T}y^1}{p^{0T}y^0}\right) \times \left(\frac{p^{1T}y^1}{p^{1T}y^0}\right)\right]^{1/2} = (Y_L \times Y_P)^{1/2}$ and $X_F = \left[\left(\frac{w^{0T}x^1}{w^{0T}x^0}\right) \times \left(\frac{w^{1T}x^1}{w^{1T}x^0}\right)\right]^{1/2} = (X_L \times X_P)^{1/2}$, subscripts L and P signifying Laspeyres and Paasche. The Malmquist quantity indexes are defined as $Y_M = \left[\left(\frac{D^0_o(x^0,y^1)}{D^0_o(x^0,y^0)}\right) \times \left(\frac{D^1_o(x^1,y^1)}{D^1_o(x^1,y^0)}\right)\right]^{1/2} = \left(Y^0_M \times Y^1_M\right)^{1/2}$ and $X_M = \left[\left(\frac{D^0_i(y^0,x^1)}{D^0_i(y^0,x^0)}\right) \times \left(\frac{D^1_i(y^1,x^1)}{D^1_i(y^1,x^0)}\right)\right]^{1/2} = \left(X^0_M \times X^1_M\right)^{1/2}$, the superscripts 0 and 1 signifying base period and comparison period, corresponding to the period weights in Laspeyres and Paasche indexes.

[5]The Fisher price indexes are defined as $P_F = \left[\left(\frac{y^{0T}p^1}{y^{0T}p^0}\right) \times \left(\frac{y^{1T}p^1}{y^{1T}p^0}\right)\right]^{1/2} = (P_L \times P_P)^{1/2}$ and $W_F = \left[\left(\frac{x^{0T}w^1}{x^{0T}w^0}\right) \times \left(\frac{x^{1T}w^1}{x^{1T}w^0}\right)\right]^{1/2} = (W_L \times W_P)^{1/2}$.

$$\begin{aligned} P_F \times Y_M\left(x^1, x^0, y^1, y^0\right) &\cong P_F \times Y_F = \frac{R^1}{R^0} \\ W_F \times X_M\left(y^1, y^0, x^1, x^0\right) &\cong W_F \times X_F = \frac{C^1}{C^0} \\ \frac{P_F}{W_F} \times \frac{Y_M(x^1, x^0, y^1, y^0)}{X_M(y^1, y^0, x^1, x^0)} &\cong \frac{P_F}{W_F} \times \frac{Y_F}{X_F} = \frac{\Pi^1}{\Pi^0}. \end{aligned} \tag{5.2}$$

Results (5.1) and (5.2) are based on Mahler inequalities, which use distance functions to bound the allocative efficiencies of quantity vectors $\{r^t(x, p) \geq p^T[y/D_o^t(x, y)]\ \forall p, y, x$ and $c^t(y, w) \leq w^T[x/D_i^t(y, x)]\ \forall w, x, y\}$, with an assumption of within-period allocative efficiency $\{p^{tT}[y^t/D_o^t(y^t, x^t)] = r^t(x^t, p^t)$ and $w^{tT}[x^t/D_i^t(x^t, y^t)] = c^t(y^t, w^t), t = 0,1\}$.

We also know that our best empirical and theoretical price and price recovery indexes are related by

$$\begin{aligned} P_F &\cong P_K(x^1, x^0, p^1, p^0) \\ W_F &\cong W_K(y^1, y^0, w^1, w^0) \\ \frac{P_F}{W_F} &\cong \frac{P_K(x^1, x^0, p^1, p^0)}{W_K(y^1, y^0, w^1, w^0)}, \end{aligned} \tag{5.3}$$

where P_F, W_F and P_F/W_F are Fisher output price, input price and price recovery indexes, and $P_K(x^1, x^0, p^1, p^0)$, $W_K(y^1, y^0, w^1, w^0)$ and $P_K(x^1, x^0, p^1, p^0)/W_K(y^1, y^0, w^1, w^0)$ are Konüs output price, input price and price recovery indexes.[6]

It follows that

$$\begin{aligned} Y_F \times P_K(x^1, x^0, p^1, p^0) &\cong Y_F \times P_F = \frac{R^1}{R^0} \\ X_F \times W_K(y^1, y^0, w^1, w^0) &\cong X_F \times W_F = \frac{C^1}{C^0} \\ \frac{Y_F}{X_F} \times \frac{P_K(x^1, x^0, p^1, p^0)}{W_K(y^1, y^0, w^1, w^0)} &\cong \frac{Y_F}{X_F} \times \frac{P_F}{W_F} = \frac{\Pi^1}{\Pi^0}. \end{aligned} \tag{5.4}$$

Results (5.3) and (5.4) are not based on Mahler inequalities. These results are based on inequalities having similar form as the Mahler inequalities $[r^t(x, p) \geq y^T p\ \forall\ y, x, p$ and $c^t(y, w) \leq x^T w\ \forall\ x, y, w]$ beneath (5.2), but they use revenue and cost frontiers to bound revenue and cost efficiencies rather than output and input

[6]The Konüs price indexes are defined as $P_K = \left[\frac{r^0(x^0,p^1)}{r^0(x^0,p^0)} \times \frac{r^1(x^1,p^1)}{r^1(x^1,p^0)}\right]^{1/2} = \left(P_K^0 \times P_K^1\right)^{1/2}$ and $W_K = \left[\frac{c^0(y^0,w^1)}{c^0(y^0,w^0)} \times \frac{c^1(y^1,w^1)}{c^1(y^1,w^0)}\right]^{1/2} = \left(W_K^0 \times W_K^1\right)^{1/2}$.

allocative efficiencies, and the efficiencies being bounded are those of output and input price vectors rather than output and input quantity vectors.

The results in (5.2) and (5.4) show that combining empirical indexes with theoretical indexes leads to approximate satisfaction of the product test with the relevant value change. It is also possible to show that combining theoretical indexes with theoretical indexes leads to approximate satisfaction of the product test with the relevant value change. Combining the left sides of (5.1) and (5.3) leads to satisfaction of the relevant product tests. Hence combining the right sides leads to approximate satisfaction of the relevant product tests and we have

$$\begin{aligned} Y_M(x^1, x^0, y^1, y^0) \times P_K(x^1, x^0, p^1, p^0) &\cong \frac{R^1}{R^0} \\ X_M(y^1, y^0, x^1, x^0) \times W_K(y^1, y^0, w^1, w^0) &\cong \frac{C^1}{C^0} \\ \frac{Y_M(x^1, x^0, y^1, y^0)}{X_M(y^1, y^0, x^1, x^0)} \times \frac{P_K(x^1, x^0, p^1, p^0)}{W_K(y^1, y^0, w^1, w^0)} &\cong \frac{\Pi^1}{\Pi^0}, \end{aligned} \tag{5.5}$$

which shows that our best theoretical quantity and productivity indexes and our best price and price recovery indexes almost satisfy the product test with the relevant value change.

In Sects. 5.3 and 5.4 we derive exact relationships between empirical and theoretical index numbers, and we provide economic interpretations of the mix functions that convert the approximations to equalities. We also show that the economic content of the quantity mix functions that convert the approximations in (5.1) and (5.2) to equalities coincide, and they differ fundamentally from the economic content of the price mix functions that convert the approximations in (5.3) and (5.4) to equalities, which also coincide. The product of the quantity mix functions and the price mix functions converts the approximations in (5.5) to equalities.

The starting points in our analyses are implicit Malmquist output and input price indexes in Sect. 5.3, and implicit Konüs output and input quantity indexes in Sect. 5.4. Neither pair of implicit indexes satisfies the fundamental homogeneity property in prices or quantities, respectively (Diewert 1981; 174, 176). However we do not treat these implicit indexes as price or quantity indexes; we use them for other purposes, to convert the economic approximations in (5.1) and (5.3) to exact relationships, which in turn eliminates the product test gaps in (5.2), (5.4) and (5.5), and to provide economic interpretations of the gaps they eliminate.

5.3 Implicit Malmquist Price and Price Recovery Indexes

In this section we exploit implicit Malmquist output price, input price and price recovery indexes. These implicit indexes enable us to derive exact relationships between Fisher and Malmquist output quantity, input quantity and productivity indexes, and exact decompositions of revenue change, cost change and profitability change.

5.3.1 The Output Side

A base period implicit Malmquist output price index is defined as

$$\begin{aligned} \mathrm{PI}_{\mathrm{M}}^{0}(\mathrm{x}^{0},\mathrm{p}^{1},\mathrm{p}^{0},\mathrm{y}^{1},\mathrm{y}^{0}) &= \frac{\mathrm{R}^{1}/\mathrm{R}^{0}}{\mathrm{Y}_{\mathrm{M}}^{0}(\mathrm{x}^{0},\mathrm{y}^{1},\mathrm{y}^{0})} \\ &= \frac{\mathrm{p}^{1\mathrm{T}}\mathrm{y}^{1}/\mathrm{D}_{\mathrm{o}}^{0}(\mathrm{x}^{0},\mathrm{y}^{1})}{\mathrm{p}^{0\mathrm{T}}\mathrm{y}^{0}/\mathrm{D}_{\mathrm{o}}^{0}(\mathrm{x}^{0},\mathrm{y}^{0})}, \end{aligned} \tag{5.6}$$

in which $\mathrm{Y}_{\mathrm{M}}^{0}(\mathrm{x}^{0},\mathrm{y}^{1},\mathrm{y}^{0}) = \mathrm{D}_{\mathrm{o}}^{0}(\mathrm{x}^{0},\mathrm{y}^{1})/\mathrm{D}_{\mathrm{o}}^{0}(\mathrm{x}^{0},\mathrm{y}^{0})$ is a base period Malmquist output quantity index. Multiplying and dividing by $\mathrm{p}^{0\mathrm{T}}\mathrm{y}^{1}/\mathrm{D}_{\mathrm{o}}^{0}(\mathrm{x}^{0},\mathrm{y}^{1})$ yields

$$\begin{aligned} \mathrm{PI}_{\mathrm{M}}^{0}(\mathrm{x}^{0},\mathrm{p}^{1},\mathrm{p}^{0},\mathrm{y}^{1},\mathrm{y}^{0}) &= \mathrm{P}_{\mathrm{P}} \times \frac{\mathrm{p}^{0\mathrm{T}}[\mathrm{y}^{1}/\mathrm{D}_{\mathrm{o}}^{0}(\mathrm{x}^{0},\mathrm{y}^{1})]}{\mathrm{p}^{0\mathrm{T}}[\mathrm{y}^{0}/\mathrm{D}_{\mathrm{o}}^{0}(\mathrm{x}^{0},\mathrm{y}^{0})]} \\ &= \mathrm{P}_{\mathrm{P}} \times \frac{\mathrm{Y}_{\mathrm{L}}}{\mathrm{Y}_{\mathrm{M}}^{0}(\mathrm{x}^{0},\mathrm{y}^{1},\mathrm{y}^{0})} \\ &= \mathrm{P}_{\mathrm{P}} \times \mathrm{YM}_{\mathrm{M}}^{0}(\mathrm{x}^{0},\mathrm{p}^{0},\mathrm{y}^{1},\mathrm{y}^{0}), \end{aligned} \tag{5.7}$$

in which $\mathrm{P}_{\mathrm{P}} = \mathrm{y}^{1\mathrm{T}}\mathrm{p}^{1}/\mathrm{y}^{1\mathrm{T}}\mathrm{p}^{0}$ is a Paasche output price index, $\mathrm{Y}_{\mathrm{L}} = \mathrm{p}^{0\mathrm{T}}\mathrm{y}^{1}/\mathrm{p}^{0\mathrm{T}}\mathrm{y}^{0}$ is a Laspeyres output quantity index, and $\mathrm{YM}_{\mathrm{M}}^{0}(\mathrm{x}^{0},\mathrm{p}^{0},\mathrm{y}^{1},\mathrm{y}^{0}) = \mathrm{p}^{0\mathrm{T}}\,[\mathrm{y}^{1}/\mathrm{D}_{\mathrm{o}}^{0}(\mathrm{x}^{0},\mathrm{y}^{1})]/\mathrm{p}^{0\mathrm{T}}\,[\mathrm{y}^{0}/\mathrm{D}_{\mathrm{o}}^{0}(\mathrm{x}^{0},\mathrm{y}^{0})]$ is a base period Malmquist output quantity mix function, so named because it is based on output distance functions that scale output vectors y^{1} and y^{0} to the base period frontier $\mathrm{IP}^{0}(\mathrm{x}^{0})$, thereby eliminating any magnitude, or technical efficiency, difference between them, leaving only difference in their mix, or allocative efficiency. This function is the ratio of the revenue generated by $\mathrm{y}^{1}/\mathrm{D}_{\mathrm{o}}^{0}(\mathrm{x}^{0},\mathrm{y}^{1})$ to that generated by $\mathrm{y}^{0}/\mathrm{D}_{\mathrm{o}}^{0}(\mathrm{x}^{0},\mathrm{y}^{0})$ when both are valued at base period output prices on base period technology. The third equality in (5.7) provides an exact decomposition of a base period implicit Malmquist output price index. The second equality demonstrates that the base period Malmquist output quantity mix function is the ratio of a Laspeyres output quantity index to a base period Malmquist output quantity index. In the presence of base period prices we expect

technically efficient base period quantities $y^0/D_o^0(x^0, y^0)$ to be at least as allocatively efficient, and therefore to generate at least as much revenue, as technically efficient comparison period quantities $y^1/D_o^0(x^0, y^1)$, and so we expect $YM_M^0(x^0, p^0, y^1, y^0) \leq 1$, and thus $Y_L \leq Y_M^0(x^0, y^1, y^0)$.

Revenue change is expressed as

$$\begin{aligned}\frac{R^1}{R^0} &\equiv Y_M^0(x^0, y^1, y^0) \times PI_M^0(x^0, p^1, p^0, y^1, y^0) \\ &= [Y_M^0(x^0, y^1, y^0) \times P_P] \times YM_M^0(x^0, p^0, y^1, y^0),\end{aligned} \tag{5.8}$$

which uses (5.7) to provide an exact decomposition of revenue change, showing that the product of a base period Malmquist output quantity index, a Paasche output price index, and a base period Malmquist output quantity mix function satisfies the product test with revenue change.

The base period output quantity mix function has a value of unity if M = 1, or if M > 1 and $y^1 = \lambda y^0$, $\lambda > 0$. If $YM_M^0(x^0, p^0, y^1, y^0) = 1$, $PI_M^0(x^0, p^1, p^0, y^1, y^0) = P_P$ and $Y_L = Y_M^0(x^0, y^1, y^0)$ in (5.7) and $R^1/R^0 = Y_M^0(x^0, y^1, y^0) \times P_P$ in (5.8), so that, under either of the stipulated conditions, a base period implicit Malmquist output price index is equal to a Paasche output price index, a base period Malmquist output quantity index is equal to a Laspeyres output quantity index, and the product of a base period Malmquist output quantity index and a Paasche output price index satisfies the product test with revenue change.

If neither of these conditions holds, we expect $YM_M^0(x^0, p^0, y^1, y^0) < 1$. Base period output allocative efficiency of $y^0/D_o^0(x^0, y^0)$ relative to p^0 is sufficient for $YM_M^0(x^0, p^0, y^1, y^0) < 1$, and thus for $PI_M^0(x^0, p^1, p^0, y^1, y^0) < P_P$, $Y_L < Y_M^0(x^0, y^1, y^0)$, and $R^1/R^0 \leq Y_M^0(x^0, y^1, y^0) \times P_P$. A less restrictive sufficient condition for all three inequalities requires only that $y^0/D_o^0(x^0, y^0)$ be more allocatively efficient than $y^1/D_o^0(x^0, y^1)$ relative to (x^0, p^0) on the frontier of base period technology $IP^0(x^0)$. This assumption is weaker than one of base period output allocative efficiency (e.g., Balk 1998) or of base period revenue maximization (e.g., Diewert 1981).

A comparison period implicit Malmquist output price index is defined as

$$\begin{aligned}PI_M^1(x^1, p^1, p^0, y^1, y^0) &\equiv \frac{R^1/R^0}{Y_M^1(x^1, y^1, y^0)} \\ &= \frac{p^{1T}[y^1/D_o^1(x^1, y^1)]}{p^{0T}[y^0/D_o^1(x^1, y^0)]},\end{aligned} \tag{5.9}$$

in which $Y_M^1(x^1, y^1, y^0) = D_o^1(x^1, y^1)/D_o^1(x^1, y^0)$ is a comparison period Malmquist output quantity index. Multiplying and dividing by $p^{1T}y^0/D_o^1(x^1, y^0)$ yields

$$\begin{aligned} PI_M^1(x^1,p^1,p^0,y^1,y^0) &= P_L \times \frac{p^{1T}[y^1/D_o^1(x^1,y^1)]}{p^{1T}[y^0/D_o^1(x^1,y^0)]} \\ &= P_L \times \frac{Y_P}{Y_M^1(x^1,y^1,y^0)}, \\ &= P_L \times YM_M^1(x^1,p^1,y^1,y^0), \end{aligned} \tag{5.10}$$

in which $P_L = y^{0T}p^1/y^{0T}p^0$ is a Laspeyres output price index, $Y_P = p^{1T}y^1/p^{1T}y^0$ is a Paasche output quantity index, and $YM_M^1(x^1,p^1,y^1,y^0) = p^{1T}\,[y^1/D_o^1(x^1,y^1)]/p^{1T}\,[y^0/D_o^1(x^1,y^0)]$ is a comparison period Malmquist output quantity mix function that is the ratio of the revenue generated by $y^1/D_o^1(x^1,y^1)$ to that generated by $y^0/D_o^1(x^1,y^0)$ when both are valued at comparison period output prices. The third equality in (5.10) provides an exact decomposition of a comparison period implicit Malmquist output price index. The second equality shows that the comparison period Malmquist output quantity mix function is the ratio of a Paasche output quantity index to a comparison period Malmquist output quantity index. In the presence of comparison period prices we expect technically efficient comparison period quantities $y^1/D_o^1(x^1,y^1)$ to be at least as allocatively efficient, and thus to generate at least as much revenue, as technically efficient base period quantities $y^0/D_o^1(x^1,y^0)$, and so we expect $YM_M^1(x^1,p^1,y^1,y^0) \geq 1$, and thus $Y_P \geq Y_M^1(x^1,y^1,y^0)$.

Revenue change is expressed as

$$\begin{aligned} \frac{R^1}{R^0} &\equiv Y_M^1(x^1,y^1,y^0) \times PI_M^1(x^1,p^1,p^0,y^1,y^0) \\ &= [Y_M^1(x^1,y^1,y^0) \times P_L] \times YM_M^1(x^1,p^1,y^1,y^0), \end{aligned} \tag{5.11}$$

which provides a second exact decomposition of revenue change, in which the product of a comparison period Malmquist output quantity index, a Laspeyres output price index, and a comparison period Malmquist output quantity mix function also satisfies the product test with revenue change.

The comparison period output quantity mix function has a value of unity if $M = 1$, or if $M > 1$ and $y^1 = \lambda y^0$, $\lambda > 0$. Under either of these conditions a comparison period implicit Malmquist output price index is equal to a Laspeyres output price index in (5.10), a comparison period Malmquist output quantity index is equal to a Paasche output quantity index in (5.10), and the product of a comparison period Malmquist output quantity index and a Laspeyres output price index satisfies the product test with revenue change in (5.11).

If neither of these conditions holds, comparison period output allocative efficiency of $y^1/D_o^1(x^1,y^1)$ relative to p^1 is sufficient for $YM_M^1(x^1,p^1,y^1,y^0) > 1$, and thus for $PI_M^1(x^1,p^1,p^0,y^1,y^0) > P_L$, $Y_P > Y_M^1(x^1,y^1,y^0)$, and $R^1/R^0 > Y_M^1(x^1,y^1,y^0) \times P_L$. A less restrictive sufficient condition for all three inequalities

requires only that $y^1/D_o^1(x^1, y^1)$ be more allocatively efficient than $y^0/D_o^1(x^1, y^0)$ relative to (x^1, p^1) on the frontier of comparison period technology $IP^1(x^1)$.

Figure 5.1 illustrates the base period and comparison period output quantity mix functions for M = 2. Convexity of the output sets, in conjunction with the condition that $y^0/D_o^0(x^0, y^0)$ be more allocatively efficient than $y^1/D_o^0(x^0, y^1)$ relative to p^0 on $IP^0(x^0)$ and that $y^1/D_o^1(x^1, y^1)$ be more allocatively efficient than $y^0/D_o^1(x^1, y^0)$ relative to p^1 on $IP^1(x^1)$, guarantees that $YM_M^0[x^0, p^0, y^1/D_o^0(x^0, y^1), y^0/D_o^0(x^0, y^0)] \leq 1$ and that $YM_M^1[x^1, p^1, y^1/D_o^1(x^1, y^1), y^0/D_o^1(x^1, y^0)] \geq 1$, and leads to the expectation that their geometric mean is approximately unity.

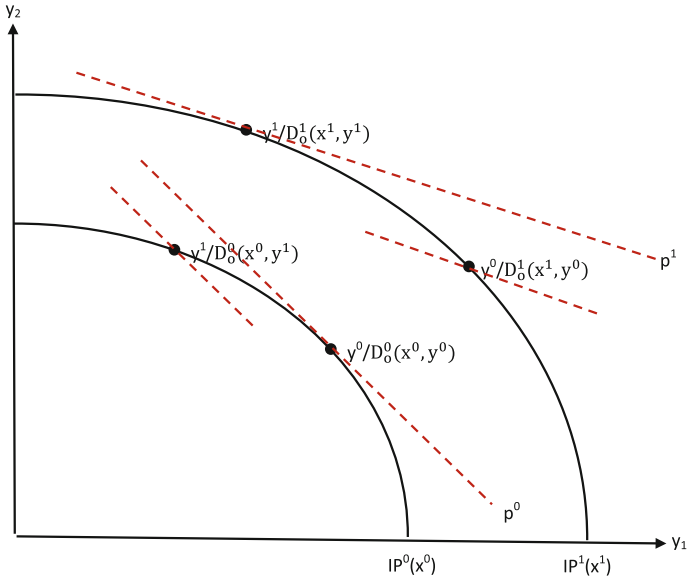


Fig. 5.1 Output quantity mix functions

An implicit Malmquist output price index is the geometric mean of (5.7) and (5.10), and so

$$\begin{aligned} PI_M(x^1, x^0, p^1, p^0, y^1, y^0) &= P_F \times YM_M(x^1, x^0, p^1, p^0, y^1, y^0) \\ &= P_F \times \frac{Y_F}{Y_M(x^1, x^0, y^1, y^0)}, \end{aligned} \tag{5.12}$$

in which $P_F = [P_P \times P_L]^{1/2}$ is a Fisher output price index, $YM_M(x^1, x^0, p^1, p^0, y^1, y^0) = [YM_M^0(x^0, p^0, y^1, y^0) \times YM_M^1(x^1, p^1, y^1, y^0)]^{1/2}$ is a Malmquist output quantity mix function, $Y_F = [Y_L \times Y_P]^{1/2}$ is a Fisher output quantity index, and $Y_M(x^1,x^0,y^1,y^0) = \left[Y_M^0(x^0, y^1, y^0) \times Y_M^1(x^1, y^1, y^0)\right]^{1/2}$ is a Malmquist output quantity index.

It follows from the first and second equalities in (5.12) that

$$Y_F = Y_M(x^1, x^0, y^1, y^0) \times YM_M(x^1, x^0, p^1, p^0, y^1, y^0), \tag{5.13}$$

which provides an exact relationship between the empirical Fisher output quantity index and the theoretical Malmquist output quantity index.

Revenue change is the geometric mean of (5.8) and (5.11), and so

$$\frac{R^1}{R^0} = [Y_M(x^1, x^0, y^1, y^0) \times P_F] \times YM_M(x^1, x^0, p^1, p^0, y^1, y^0), \tag{5.14}$$

which provides an exact decomposition of revenue change.

Summarizing the output side, using (5.7) and (5.10), the Malmquist output quantity mix function is given by

$$YM_M(x^1, x^0, p^1, p^0, y^1, y^0) = \left\{ \frac{p^{0T}[\frac{y^1}{D_o^0(x^0, y^1)}]}{p^{0T}[\frac{y^0}{D_o^0(x^0, y^0)}]} \times \frac{p^{1T}[\frac{y^1}{D_o^1(x^1, y^1)}]}{p^{1T}[\frac{y^0}{D_o^1(x^1, y^0)}]} \right\}^{1/2}, \tag{5.15}$$

and has a value of unity if M = 1 or if $y^1 = \lambda y^0$, $\lambda > 0$. Otherwise we expect it to be approximately unity, which leads to the economically meaningful expectations that $Y_F \cong Y_M(x^1, x^0, y^1, y^0)$ in (5.13) and $R^1/R^0 \cong Y_M(x^1, x^0, y^1, y^0) \times P_F$ in (5.14).[7]

5.3.2 *The Input Side*

We exploit the implicit Malmquist input price index in a similar manner, using the same strategies and the same quantity mix logic. The base period implicit Malmquist input price index is $WI_M^0(y^0, w^1, w^0, x^1, x^0) \equiv (C^1/C^0)/X_M^0(y^0, x^1, x^0)$ and the comparison period implicit Malmquist input price index is $WI_M^1(y^1, w^1, w^0, x^1, x^0) \equiv (C^1/C^0)/X_M^1(y^1, x^1, x^0)$. We omit all intermediate steps and arrive at the geometric mean of the two, the implicit Malmquist input price index

[7]The fact that $YM_M(x^1, x^0, p^1, p^0, y^1, y^0) = \left[YM_M^0(x^0, p^0, y^1, y^0) \times YM_M^1(x^1, p^1, y^1, y^0)\right]^{1/2} = [(\leq 1) \times (\geq 1)]^{1/2}$ leads to the expectation that $YM_M(x^1, x^0, p^1, p^0, y^1, y^0) \cong 1$, but it does not guarantee that the approximation be numerically close. Balk (1998; 37) provides the mathematical reasoning; the economic reasoning requires that the allocative efficiency ratios be roughly reciprocal in base and comparison periods. However if allocative efficiency improves or declines sufficiently from base to comparison periods, the approximation may not be numerically close. This qualification to the economic argument for closeness of a geometric mean approximation also applies to (5.19) on the input side, and to (5.28) and (5.35) in Sect. 5.4.

$$\begin{aligned} WI_M(y^1, y^0, w^1, w^0, x^1, x^0) &= W_F \times \left[\frac{w^{0T}[x^1/D_i^0(y^0, x^1)]}{w^{0T}[x^0/D_i^0(y^0, x^0)]} \times \frac{w^{1T}[x^1/D_i^1(y^1, x^1)]}{w^{1T}[x^0/D_i^1(y^1, x^0)]} \right]^{1/2} \\ &= W_F \times \frac{X_F}{X_M(y^1, y^0, x^1, x^0)} \\ &= W_F \times XM_M(y^1, y^0, w^1, w^0, x^1, x^0), \end{aligned} \tag{5.16}$$

in which the Fisher input price index $W_F = [W_P \times W_L]^{1/2}$, the Fisher input quantity index $X_F = [X_L \times X_P]^{1/2}$, and the Malmquist input quantity index $X_M(y^1, y^0, x^1, x^0) = [X_M^0(y^0, x^1, x^0) \times X_M^1(y^1, x^1, x^0)]^{1/2}$. The Malmquist input quantity mix function $XM_M(y^1, y^0, w^1, w^0, x^1, x^0)$ is the geometric mean of a base period Malmquist input quantity mix function that is the ratio of the cost incurred at $x^1/D_i^0(y^0, x^1)$ to that at $x^0/D_i^0(y^0, x^0)$ when both are valued at base period input prices on the frontier of base period technology $IL^0(y^0)$, and a comparison period Malmquist input quantity mix function that is the ratio of the cost incurred at $x^1/D_i^1(y^1, x^1)$ to that at $x^0/D_i^1(y^1, x^0)$ when both are valued at comparison period input prices on the frontier of comparison period technology $IL^1(y^1)$. The third equality in (5.16) provides an exact decomposition of the implicit Malmquist input price index. The second equality shows that the Malmquist input quantity mix function is the ratio of a Fisher input quantity index to a Malmquist input quantity index, from which it follows that

$$X_F = X_M(y^1, y^0, x^1, x^0) \times XM_M(y^1, y^0, w^1, w^0, x^1, x^0), \tag{5.17}$$

which provides an exact relationship between an empirical Fisher input quantity index and a theoretical Malmquist input quantity index.

Since cost change can be expressed as $C^1/C^0 = X_F \times W_F$, it follows from (5.17) that

$$\frac{C^1}{C^0} = [X_M(y^1, y^0, x^1, x^0) \times W_F] \times XM_M(y^1, y^0, w^1, w^0, x^1, x^0), \tag{5.18}$$

which provides an exact decomposition of cost change.

The input quantity mix function has a value of unity if $N = 1$, or if $x^1 = \mu x^0$, $\mu > 0$. Under either of these conditions $WI_M(y^1, y^0, w^1, w^0, x^1, x^0) = W_F$ in (5.16), $X_F = X_M(y^1, y^0, x^1, x^0)$ in (5.17), and $C^1/C^0 = X_M(y^1, y^0, x^1, x^0) \times W_F$ in (5.18). If neither of these conditions holds, we expect $XM_M(y^1, y^0, w^1, w^0, x^1, x^0) \cong 1$, even in the absence of within-period input allocative efficiency, which generates $WI_M(y^1, y^0, w^1, w^0, x^1, x^0) \cong W_F$ in (5.16), $X_F \cong X_M(y^1, y^0, x^1, x^0)$ in (5.17), and $C^1/C^0 \cong X_M(y^1, y^0, x^1, x^0) \times W_F$ in (5.18).

Figure 5.2 illustrates the base period and comparison period input quantity mix functions with N = 2. Convexity of the input sets, in conjunction with the condition that $x^0/D_i^0(y^0, x^0)$ be more allocatively efficient than $x^1/D_i^0(y^0, x^1)$ relative to w^0 on $IL^0(y^0)$ and that $x^1/D_i^1(y^1, x^1)$ be more allocatively efficient than $x^0/D_i^1(y^1, x^0)$ relative to w^1 on $IL^1(y^1)$, guarantees that $XM_M^0[y^0, w^1, w^0, x^1/D_i^0(y^0, x^1), x^0/D_i^0(y^0, x^0)] \geq 1$ and that $XM_M^1[y^1, w^1, w^0, x^1/D_i^1(y^1, x^1), x^0/D_i^1(y^1, x^0)] \leq 1$, and leads to the expectation that their geometric mean is approximately unity.

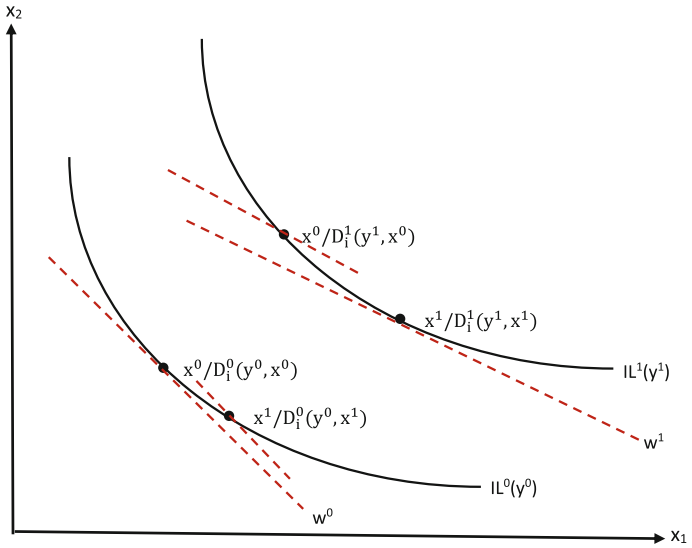


Fig. 5.2 Input quantity mix functions

Summarizing the input side, from (5.16) the Malmquist input quantity mix function is given by

$$XM_M(y^1, y^0, w^1, w^0, x^1, x^0) = \left\{ \frac{w^{0T}[x^1/D_i^0(y^0, x^1)]}{w^{0T}[x^0/D_i^0(y^0, x^0)]} \times \frac{w^{1T}[x^1/D_i^1(y^1, x^1)]}{w^{1T}[x^0/D_i^1(y^1, x^0)]} \right\}^{1/2}, \tag{5.19}$$

and has a value of unity if N = 1 or if $x^1 = \mu x^0$, $\mu > 0$. Otherwise we expect it to be approximately unity, which leads to the economically meaningful expectations that $X_F \cong X_M(y^1, y^0, x^1, x^0)$ in (5.17) and $C^1/C^0 \cong X_M(y^1, y^0, x^1, x^0) \times W_F$ in (5.18).

5.3.3 Combining the Output Side and the Input Side

We ignore base period and comparison period indexes and proceed directly to an implicit Malmquist price recovery index. The ratio of (5.12) and (5.16) is

$$\frac{\mathrm{PI_M}(x^1,x^0,p^1,p^0,y^1,y^0)}{\mathrm{WI_M}(y^1,y^0,w^1,w^0,x^1,x^0)}=\frac{P_F}{W_F}\times \mathrm{M_M}(y^1,y^0,x^1,x^0,p^1,p^0,w^1,w^0), \tag{5.20}$$

in which $\mathrm{M_M}(y^1,y^0,x^1,x^0,p^1,p^0,w^1,w^0) = \mathrm{YM_M}(x^1,x^0,p^1,p^0,y^1,y^0)/\mathrm{XM_M}(y^1,y^0,w^1,w^0,x^1,x^0)$ is a Malmquist quantity mix function that provides an economic characterization of the gap, if any, between P_F/W_F and $\mathrm{PI_M}(x^1,x^0,p^1,p^0,y^1,y^0)/\mathrm{WI_M}(y^1,y^0,w^1,w^0,x^1,x^0)$. From (5.12) to (5.14) we expect $\mathrm{YM_M}(x^1,x^0,p^1,p^0,y^1,y^0) \cong 1$, and from (5.16) to (5.18) we expect $\mathrm{XM_M}(y^1,y^0,w^1,w^0,x^1,x^0) \cong 1$. Consequently we expect $\mathrm{M_M}(y^1,y^0,x^1,x^0,p^1,p^0,w^1,w^0) \cong 1$, in which case a Fisher price recovery index is approximately equal to an implicit Malmquist price recovery index.

A geometric mean expression for productivity change is given by the ratio of (5.13) and (5.17), and is

$$\frac{Y_F}{X_F}=\frac{Y_M(x^1,x^0,y^1,y^0)}{X_M(y^1,y^0,x^1,x^0)}\times \mathrm{M_M}(y^1,y^0,x^1,x^0,p^1,p^0,w^1,w^0), \tag{5.21}$$

which provides an exact relationship between a Fisher productivity index and a Malmquist productivity index, with the Malmquist quantity mix function providing an economic interpretation of the (presumably small) gap between the two.

A geometric mean expression for profitability change is given by the ratio of (5.14) and (5.18), and is

$$\frac{\Pi^1}{\Pi^0}=\left[\frac{Y_M(x^1,x^0,y^1,y^0)}{X_M(y^1,y^0,x^1,x^0)}\times\frac{P_F}{W_F}\right]\times \mathrm{M_M}(y^1,y^0,x^1,x^0,p^1,p^0,w^1,w^0), \tag{5.22}$$

which provides an exact decomposition of profitability change. If the Malmquist quantity mix function is approximately unity a Malmquist productivity index and a Fisher price recovery index approximately satisfy the product test with profitability change.[8]

In this section we have used implicit Malmquist price and price recovery indexes to relate empirical Fisher quantity and productivity indexes to theoretical Malmquist quantity and productivity indexes. The important findings are contained in (5.12), (5.16) and (5.20); (5.13), (5.17) and (5.21); and (5.14), (5.18) and (5.22).

[8]All approximation results in this section also can occur if the technologies allow infinite output substitution possibilities between output rays defined by y^1 and y^0 along $IP^0(x^0)$ and $IP^1(x^1)$ in Fig. 1, and infinite input substitution possibilities between input rays defined by x^1 and x^0 along $IL^0(y^0)$ and $IL^1(y^1)$ in Fig. 2. DEA generates such technologies.

The first set of results relates implicit theoretical price and price recovery indexes to their explicit empirical counterparts, and establishes the foundations for the second and third sets of results. Equations (5.13), (5.17) and (5.21) clarify the sense in which Fisher quantity and productivity indexes and Malmquist quantity and productivity indexes are approximately equal. Equations (5.14), (5.18) and (5.22) clarify the sense in which Malmquist quantity and productivity indexes approximately satisfy the relevant product test with Fisher price and price recovery indexes. Both sets of results depends fundamentally on Malmquist output and input quantity mix functions, which have clear economic interpretations. It is worth emphasizing that the quantity mix functions compare the allocative efficiencies of pairs of technically efficient *quantity* vectors, which are the choice variables in the exercises.[9]

Equations (5.14), (5.18) and (5.22) warrant special emphasis from an empirical perspective, because of their decomposability properties. $Y_M(x^1, x^0, y^1, y^0)$, $X_M(y^1, y^0, x^1, x^0)$ and $Y_M(x^1, x^0, y^1, y^0)/X_M(y^1, y^0, x^1, x^0)$ decompose into the product of economic drivers of productivity change: technical change, technical efficiency change, mix efficiency change and size efficiency change. In contrast, P_F, W_F and P_F/W_F decompose into contributions of individual output and input price changes (Balk (2004)). These two features enable a decomposition of value (revenue, cost and profitability) change into the economic drivers of quantity change and the individual price drivers of price change.

5.4 Implicit Konüs Quantity and Productivity Indexes

In this section we exploit implicit Konüs output quantity, input quantity and productivity indexes. These implicit indexes lead us to exact relationships between Fisher and Konüs output price, input price and price recovery indexes, and to exact decompositions of revenue change, cost change and profitability change. Both sets of results differ from analogous results in Sect. 5.3.

5.4.1 The Output Side

We begin with a base period implicit Konüs output quantity index, which is defined as

[9]The quantity mix functions in Sect. 5.3 are ratio analogues to the product mix and resource mix effects in Grifell-Tatjé and Lovell (1999; 1182, 1184).

$$\begin{aligned} YI_K^0(x^0, p^1, p^0, y^1, y^0) &\equiv \frac{R^1/R^0}{P_K^0(x^0, p^1, p^0)} \\ &= \frac{y^{1T}[p^1/r^0(x^0, p^1)]}{y^{0T}[p^0/r^0(x^0, p^0)]}, \end{aligned} \tag{5.23}$$

in which $P_K^0(x^0, p^1, p^0) = r^0(x^0, p^1)/r^0(x^0, p^0)$ is a base period Konüs output price index. Multiplying and dividing by $y^{0T}p^1/r^0(x^0, p^1)$ yields

$$\begin{aligned} YI_K^0(x^0, p^1, p^0, y^1, y^0) &= Y_P \times \frac{y^{0T}[p^1/r^0(x^0, p^1)]}{y^{0T}[p^0/r^0(x^0, p^0)]} \\ &= Y_P \times \frac{P_L}{P_K^0(x^0, p^1, p^0)} \\ &= Y_P \times PM_K^0(x^0, y^0, p^1, p^0), \end{aligned} \tag{5.24}$$

in which $Y_P = p^{1T}y^1/p^{1T}y^0$ is a Paasche output quantity index, $P_L = y^{0T}p^1/y^{0T}p^0$ is a Laspeyres output price index, and $PM_K^0(x^0, y^0, p^1, p^0) = y^{0T}[p^1/r^0(x^0, p^1)]/y^{0T}[p^0/r^0(x^0, p^0)]$ is a base period Konüs output price mix function, so named because it is a function of revenue functions that coincide apart from their output price vectors. This function is the ratio of the revenue generated at y^0 by normalized comparison period output prices $p^1/r^0(x^0, p^1)$ to that generated at y^0 by normalized base period output prices $p^0/r^0(x^0, p^0)$. The two normalized price vectors differ only in their output price mix.[10]

The third equality in (5.24) provides an exact decomposition of a base period implicit Konüs output quantity index. The second equality demonstrates that the base period Konüs output price mix function is the ratio of a Laspeyres output price index to a base period Konüs output price index. This mix function is bounded above by unity if $p^0/r^0(x^0, p^0)$ is revenue efficient at $y^0 \in IP^0(x^0)$, or if $p^0/r^0(x^0, p^0)$ is more revenue efficient than $p^1/r^0(x^0, p^1)$ at $y^0 \in IP^0(x^0)$. In either case $P_L \leq P_K^0(x^0, p^1, p^0)$ and $YI_K^0(x^0, p^1, p^0, y^1, y^0) \leq Y_P$. $YI_K^0(x^0, p^1, p^0, y^1, y^0) = Y_P$ if either $M = 1$ or $p^1 = \lambda p^0$, $\lambda > 0$. These bounds do not require base period revenue maximizing behavior, or even base period allocative efficiency.

Revenue change is expressed as

[10] An alternative interpretation of the base period Konüs output price mix function is that it is the ratio of two revenue efficiencies, both with base period technology and quantity vectors but with different output price vectors, since it can be expressed

$$PM_K^0(x^0, y^0, p^1, p^0) = p^{1T}[y^0/r^0(x^0, p^1)]/p^{0T}[y^0/r^0(x^0, p^0)].$$

.

$$\begin{aligned}\frac{R^1}{R^0} &= P_K^0(x^0, p^1, p^0) \times YI_K^0(x^0, p^1, p^0, y^1, y^0) \\ &= [P_K^0(x^0, p^1, p^0) \times Y_P] \times PM_K^0(x^0, y^0, p^1, p^0),\end{aligned} \tag{5.25}$$

which states that the product of a base period Konüs output price index, a Paasche output quantity index and a base period Konüs output price mix function satisfies the product test with R^1/R^0. As above we expect $R^1/R^0 \leq P_K^0(x^0, p^1, p^0) \times Y_P$. However if either $M = 1$ or $p^1 = \lambda p^0$, $\lambda > 0$, (5.24) and (5.25) collapse to $YI_K^0(x^0, p^1, p^0, y^1, y^0) = Y_P$ and $R^1/R^0 = P_K^0(x^0, p^1, p^0) \times Y_P$ in which case a base period implicit Konüs output quantity index is equal to a Paasche output quantity index, and consequently a Konüs output price index and a Paasche output quantity index satisfy the product test with R^1/R^0.

We now sketch the results of a comparison period implicit Konüs output quantity index. Following the same procedures as above, after multiplying and dividing by $y^{1T}p^0/r^1(x^1, p^0)$ we have

$$\begin{aligned}YI_K^1(x^1, p^1, p^0, y^1, y^0) &= \frac{R^1/R^0}{P_K^1(x^1, p^1, p^0)} \\ &= Y_L \times \frac{y^{1T}[p^1/r^1(x^1, p^1)]}{y^{1T}[p^0/r^1(x^1, p^0)]} \\ &= Y_L \times \frac{P_P}{P_K^1(x^1, p^1, p^0)} \\ &= Y_L \times PM_K^1(x^1, y^1, p^1, p^0),\end{aligned} \tag{5.26}$$

in which $Y_L = p^{0T}y^1/p^{0T}y^0$ is a Laspeyres output quantity index, $P_P = y^{1T}p^1/y^{1T}p^0$ is a Paasche output price index, and $P_K^1(x^1, p^1, p^0) = r^1(x^1, p^1)/r^1(x^1, p^0)$ is a comparison period Konüs output price index. The comparison period Konüs output price mix function $PM_K^1(x^1, y^1, p^1, p^0)$ is the ratio of the revenue efficiency of two normalized output price vectors, given comparison period technology and quantity vectors. If, as expected, $p^1/r^1(x^1, p^1)$ is more revenue efficient than $p^0/r^1(x^1, p^0)$ at $y^1 \in IP^1(x^1)$, then $PM_K^1(x^1, y^1, p^1, p^0) \geq 1$, $YI_K^1(x^1, y^1, y^0) \geq Y_L$ and $P_P \geq P_K^1(x^1, p^1, p^0)$.

Revenue change is expressed as

$$\begin{aligned}\frac{R^1}{R^0} &= P_K^1(x^1, p^1, p^0) \times YI_K^1(x^1, p^1, p^0, y^1, y^0) \\ &= [P_K^1(x^1, p^1, p^0) \times Y_L] \times PM_K^1(x^1, y^1, p^1, p^0),\end{aligned} \tag{5.27}$$

which states that the product of a comparison period Konüs output price index, a Laspeyres output quantity index and a comparison period Konüs output price mix function satisfies the product test with R^1/R^0. Under the conditions specified above, we expect $R^1/R^0 \geq P_K^1(x^1, p^1, p^0) \times Y_L$.

Figure 5.3 illustrates the base period and comparison period output price mix functions for M = 2. Convexity of the output sets, together with the conditions that $p^0/r^0(x^0, p^0)$ be more revenue efficient than $p^1/r^0(x^0, p^1)$ at $y^0 \in IP^0(x^0)$ and that $p^1/r^1(x^1, p^1)$ be more revenue efficient than $p^0/r^1(x^1, p^0)$ at $y^1 \in IP^1(x^1)$ guarantees that $PM_K^0[x^0, y^0, p^1/r^0(x^0, p^1), p^0/r^0(x^0, p^0)] \leq 1$ and $PM_K^1[x^1, y^1, p^1/r^1(x^1, p^1), p^0/r^1(x^1, p^0)] \geq 1$.

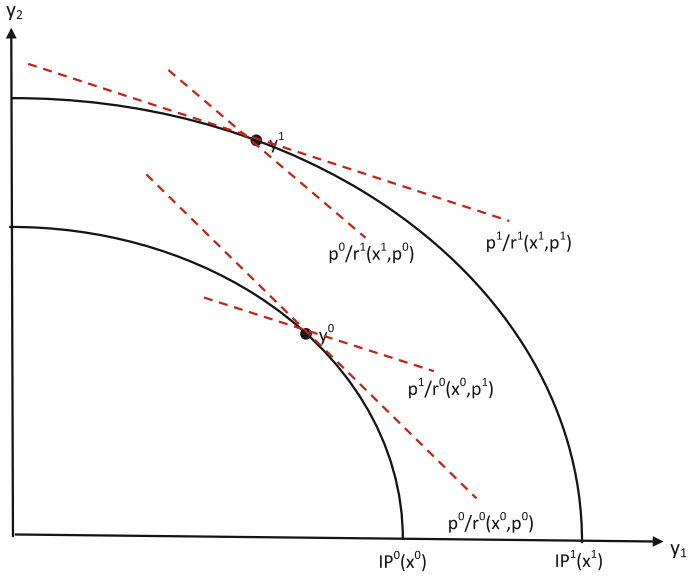


Fig. 5.3 Output price mix functions

The geometric mean of (5.24) and (5.26) is an implicit Konüs output quantity index

$$\begin{aligned} YI_K(x^1, x^0, p^1, p^0, y^1, y^0) &= Y_F \times \left[PM_K^0(x^0, y^0, p^1, p^0) \times PM_K^1(x^1, y^1, p^1, p^0)\right]^{1/2} \\ &= Y_F \times \frac{P_F}{P_K(x^1, x^0, p^1, p^0)} \\ &= Y_F \times PM_K(x^1, x^0, y^1, y^0, p^1, p^0), \end{aligned} \tag{5.28}$$

which states that an implicit Konüs output quantity index is the product of a Fisher output quantity index and a Konüs output price mix function. Because one component of the output price mix function is bounded above by unity and the other is bounded below by unity we expect $YI_K(x^1, x^0, p^1, p^0, y^1, y^0) \approx Y_F$.

It follows from the second and third equalities in (5.28) that

$$P_F = P_K(x^1, x^0, p^1, p^0) \times PM_K(x^1, x^0, y^1, y^0, p^1, p^0), \tag{5.29}$$

which enables us to calculate, and provide an economic interpretation of, the gap between the theoretical and empirical output price indexes.

The geometric mean of (5.25) and (5.27) yields the expression for revenue change

$$\frac{R^1}{R^0} = [P_K(x^1, x^0, p^1, p^0) \times Y_F] \times PM_K(x^1, x^0, y^1, y^0, p^1, p^0), \tag{5.30}$$

which provides an exact decomposition of revenue change, in which the product of a Konüs output price index, a Fisher output quantity index, and a Konüs output price mix function satisfies the product test with revenue change, the approximation becoming an equality if either $M = 1$ or $p^1 = \lambda p^0$, $\lambda > 0$.

Summarizing the output side, the Konüs output price mix function

$$PM_K(x^1, x^0, y^1, y^0, p^1, p^0) = \left\{ \frac{y^{0T}\left[\frac{p^1}{r^0(x^0,p^1)}\right]}{y^{0T}\left[\frac{p^0}{r^0(x^0,p^0)}\right]} \times \frac{y^{1T}\left[\frac{p^1}{r^1(x^1,p^1)}\right]}{y^{1T}\left[\frac{p^0}{r^1(x^1,p^0)}\right]} \right\}^{1/2} \tag{5.31}$$

has an approximate unitary value, which generates the expectation that $P_F \cong P_K(x^1, x^0, p^1, p^0)$ in (5.29) and $\frac{R^1}{R^0} \cong P_K(x^1, x^0, p^1, p^0) \times Y_F$ in (5.30).

5.4.2 The Input Side

We now consider the implicit Konüs input quantity index. The base period implicit Konüs input quantity index is $XI_K^0(y^0, w^1, w^0, x^1, x^0) = (C^1/C^0)/W_K^0(y^0, w^1, w^0)$ and the comparison period implicit Konüs input quantity index is $XI_K^1(y^1, w^1, w^0, x^1, x^0) = (C^1/C^0)/W_K^1(y^1, w^1, w^0)$. The geometric mean of the two, the implicit Konüs input quantity index, is

$$\begin{aligned} XI_K(y^1, y^0, w^1, w^0, x^1, x^0) &= X_F \times \left[WM_K^0(y^0, x^0, w^1, w^0) \times WM_K^1(y^1, x^1, w^1, w^0)\right]^{1/2} \\ &= X_F \times \frac{W_F}{W_K(y^1, y^0, w^1, w^0)} \\ &= X_F \times WM_K(y^1, y^0, x^1, x^0, w^1, w^0), \end{aligned} \tag{5.32}$$

in which the Konüs input price mix function $WM_K(y^1, y^0, x^1, x^0, w^1, w^0)$ measures the gap between $XI_K(y^1, y^0, w^1, w^0, x^1, x^0)$ and X_F, and is defined analogously to the output price mix function in (5.28).

From the second and third equalities in (5.32)

$$W_F = W_K(y^1, y^0, w^1, w^0) \times WM_K\left(y^1, y^0, x^1, x^0, w^1, w^0\right), \tag{5.33}$$

which provides an exact relationship between empirical Fisher and theoretical Konüs input price indexes.

Cost change is

$$\frac{C^1}{C^0} = [W_K(y^1, y^0, w^1, w^0) \times X_F] \times WM_K\left(y^1, y^0, x^1, x^0, w^1, w^0\right). \tag{5.34}$$

The base period and comparison period input price mix functions are illustrated in Fig. 5.4 for N = 2. Convexity of the input sets, together with the conditions that $w^0/c^0(y^0, w^0)$ be more cost efficient than $w^1/c^0(y^0, w^1)$ at $x^0 \in IL^0(y^0)$ and that $w^1/c^1(y^1, w^1)$ be more cost efficient than $w^0/c^1(y^1, w^0)$ at $x^1 \in IL^1(y^1)$ guarantees that $WM_K^0[y^0, x^0, w^1/c^0(y^0, w^1), w^0/c^0(y^0, w^0)] \geq 1$ and $WM_K^1[y^1, x^1, w^1/c^1(y^1, w^1), w^0/c^1(y^1, w^0)] \leq 1$.

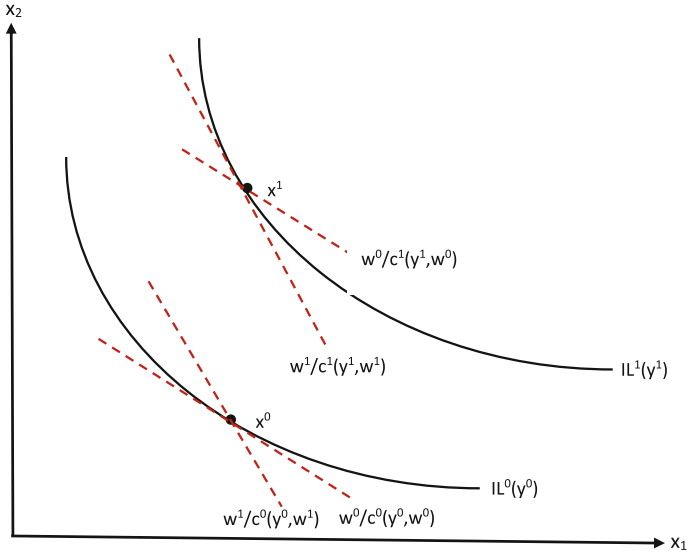


Fig. 5.4 Input price mix functions

Summarizing the input side, the Konüs input price mix function

$$WM_K\left(y^1, y^0, x^1, x^0, w^1, w^0\right) = \left\{ \frac{x^{0T}\left[\frac{w^1}{c^0\left(y^0, w^1\right)}\right]}{x^{0T}\left[\frac{w^0}{c^0\left(y^0, w^0\right)}\right]} \times \frac{x^{1T}\left[\frac{w^1}{c^1\left(y^1, w^1\right)}\right]}{x^{1T}\left[\frac{w^0}{c^1\left(y^1, w^0\right)}\right]} \right\}^{1/2} \tag{5.35}$$

has an approximate unitary value, which generates the expectation that $W_F \cong W_K(y^1, y^0, w^1, w^0)$ in (5.33) and $\frac{C^1}{C^0} \cong W_K(y^1, y^0, w^1, w^0) \times X_F$ in (5.34).

5.4.3 Combining the Output Side and the Input Side

We now construct an implicit Konüs productivity index. We ignore base period and comparison indexes and proceed directly to an implicit Konüs productivity index. The ratio of (5.28) and (5.32) is

$$\frac{YI_K(x^1, x^0, p^1, p^0, y^1, y^0)}{XI_K(y^1, y^0, w^1, w^0, x^1, x^0)} = \frac{Y_F}{X_F} \times M_K(y^1, y^0, x^1, x^0, p^1, p^0, w^1, w^0), \tag{5.36}$$

in which the Konüs price mix function $M_K(y^1, y^0, x^1, x^0, p^1, p^0, w^1, w^0) = PM_K(x^1, x^0, y^1, y^0, p^1, p^0)/WM_K(y^1, y^0, x^1, x^0, w^1, w^0)$ measures, and provides an economic interpretation of, the gap between $YI_K(x^1, x^0, p^1, p^0, y^1, y^0)/XI_K(y^1, y^0, w^1, w^0, x^1, x^0)$ and Y_F/X_F. Because we expect $PM_K(x^1, x^0, y^1, y^0, p^1, p^0) \approx 1$ and we expect $WM_K(y^1, y^0, x^1, x^0, w^1, w^0) \approx 1$ we also expect $M_K(y^1, y^0, x^1, x^0, p^1, p^0, w^1, w^0) \approx 1$, in which case the implicit Konüs productivity index is approximately equal to a Fisher productivity index.

The ratio of (5.29) and (5.33)

$$\frac{P_F}{W_F} = \frac{P_K(x^1, x^0, p^1, p^0)}{W_K(y^1, y^0, w^1, w^0)} \times M_K(y^1, y^0, x^1, x^0, p^1, p^0, w^1, w^0), \tag{5.37}$$

provides an exact relationship between an empirical Fisher price recovery index and a theoretical Konüs price recovery index.

The ratio of (5.30) and (5.34) provides an implicit Konüs measure of profitability change

$$\begin{aligned} \frac{\Pi^1}{\Pi^0} &= \frac{R^1/R^0}{C^1/C^0} \\ &= \left[\frac{P_K(x^1, x^0, p^1, p^0)}{W_K(y^1, y^0, w^1, w^0)} \times \frac{Y_F}{X_F}\right] \times M_K(y^1, y^0, x^1, x^0, p^1, p^0, w^1, w^0), \end{aligned} \tag{5.38}$$

and if $M_K(y^1, y^0, x^1, x^0, p^1, p^0, w^1, w^0) \approx 1$ a Konüs price recovery index and a Fisher productivity index approximately satisfy the product test with profitability change.[11]

[11] All approximation results in this section also can occur if y^0 and y^1 in Fig. 3 and x^0 and x^1 in Fig. 4 are vertices of piecewise linear technologies that allow $p^{1T}y^0 = p^{0T}y^0$, $p^{0T}y^1 = p^{1T}y^1$ and $w^{1T}x^0 = w^{0T}x^0$, $w^{0T}x^1 = w^{1T}x^1$, as might occur with DEA.

In this section we have used implicit Konüs quantity and productivity indexes to relate empirical Fisher price and price recovery indexes to theoretical Konüs price and price recovery indexes. The important findings are contained in (5.28), (5.32) and (5.36); (5.29), (5.33) and (5.37); and (5.30), (5.34) and (5.38). The first three relate implicit theoretical quantity and productivity indexes to their explicit empirical counterparts, and establish the foundations for the second and third sets of results. Equations (5.29), (5.33) and (5.37) clarify the sense in which Konüs price and price recovery indexes approximate Fisher price and price recovery indexes. Equations (5.30), (5.34) and (5.38) clarify the sense in which Konüs price and price recovery indexes approximately satisfy the relevant product test with Fisher quantity and productivity indexes. In both the second and third sets of results clarity is provided by the relevant Konüs price mix function.

In Sect. 5.3 the product test expressions (5.14), (5.18) and (5.22) have useful empirical applications, since Malmquist quantity and productivity indexes decompose by economic driver and Fisher price and price recovery indexes decompose by individual prices. Here the product test expressions (5.30), (5.34) and (5.38) are of potential, but as yet unrealized, empirical value. The Fisher quantity and productivity indexes have been decomposed by economic driver of productivity change, although agreement on a preferred decomposition remains elusive.[12] The Konüs price and price recovery indexes have yet to be decomposed by economic drivers of price change (rather than, as commonly practiced, by individual prices), although as we noted in Sect. 5.1 Konüs indexes are multiplicatively complete, and research on this issue is underway.

We emphasize that the Konüs price mix functions differ significantly from the Malmquist quantity mix functions, although they serve the same purposes, to convert approximations to exact relationships and to close product test gaps. The Malmquist quantity mix functions are ratios of values generated by two normalized quantity vectors weighted by a common price vector, with choice variables being quantities. The Konüs price mix functions are ratios of values generated by a single quantity vector weighted by two normalized price vectors, with choice variables being prices.

5.5 Summary and Conclusions

We have exploited implicit Malmquist price and price recovery indexes to derive exact relationships between empirical Fisher and theoretical Malmquist quantity and productivity indexes, and to derive economically meaningful functions describing the ability of Malmquist quantity and productivity indexes to satisfy

[12]Compare, for example, the decompositions proposed by Ray and Mukherjee (1996) and by Kuosmanen and Sipiläinen (2009), and the critiques of both approaches by Diewert (2014) and Grifell-Tatjé and Lovell (2015; 125–35).

product tests with Fisher price and price recovery indexes. The key to these exact relationships is the concept of Malmquist output and input quantity mix functions, in which quantities are allowed to vary between base and comparison periods but prices are fixed at either base period values or comparison period values. The important empirical implication of our analysis is that, as the variation in the quantity mix between base and comparison periods narrows (expands), the gap between Fisher and Malmquist quantity and productivity indexes also narrows (expands).

We also have exploited implicit Konüs quantity and productivity indexes to derive exact relationships between empirical Fisher and theoretical Konüs price and price recovery indexes, and to derive fundamentally different, but nonetheless economically meaningful functions describing the ability of Konüs price and price recovery indexes to satisfy product tests with Fisher quantity and productivity indexes. The key to these exact relationships is the concept of Konüs output and input price mix functions, in which prices are allowed to vary between base and comparison periods but quantities are fixed at either base period values or comparison period values. In this case, as the variation in the price mix between base and comparison periods narrows (expands), the gap between Fisher and Konüs price and price recovery indexes also narrows (expands).

The exact relationships have clear economic interpretations, as allocative efficiency effects, although these effects differ between Sects. 5.3 and 5.4. These allocative efficiency effects are easy to calculate, using data required to calculate Fisher indexes and estimate Malmquist and Konüs indexes, as Brea, Grifell-Tatjé and Lovell (2011) have demonstrated for Fisher/Malmquist pairings.[13]

Acknowledgements We are grateful to Hideyuki Mizobuchi for his very helpful comments on a previous draft. We have also benefitted from comments we have received at seminars at the 2015 International Workshop on Efficiency and Productivity held in Alicante, Spain, June 12 & 13, at the 2015 Taiwan Productivity and Efficiency Conference held at Soochow University, Taipei, Taiwan, July 7–9, and at a productivity workshop hosted by the Centre for Efficiency and Productivity Analysis at the University of Queensland, Brisbane, Australia, October 1, 2015. E. Grifell-Tatjé acknowledges the financial support from the Spanish Ministry of Science and Innovation (Ref. ECO2013-46954-C3-2-R).

References

Balk BM (1998) Industrial price, quantity and productivity indexes. Kluwer Academic Publishers, Boston

Balk BM (2004) Decompositions of fisher indexes. Econ Lett 82:107–113

Bjurek H (1996) The Malmquist total factor productivity index. Scand J Econ 98:303–313

[13]Törnqvist indexes also provide good approximations to Malmquist and Konüs indexes, and so our findings linking Fisher indexes with Malmquist and Konüs indexes might provide insight into the relationship between Törnqvist and Fisher indexes, a relationship recently explored by Mizobuchi (2016).

Brea H, Grifell-Tatjé E, Lovell CAK (2011) Testing the product test. Econ Lett 113:157–159
Caves DW, Christensen LR, Diewert WE (1982) The economic theory of index numbers and the measurement of input, output, and productivity. Econometrica 50:1393–1414
Diewert WE (1981) The economic theory of index numbers: a survey. In: Deaton A (ed) Essays in the theory and measurement of consumer behaviour in honour of sir richard stone. Cambridge University Press, Cambridge (Chapter 7)
Diewert WE (1992) Fisher ideal output, input and productivity indexes revisited. J Prod Anal 3:211–248
Diewert WE (2014) Decompositions of profitability change using cost functions. J Econometrics 183:58–66
Grifell-Tatjé E, Lovell CAK (1999) Profits and productivity. Manage Sci 45:1177–1193
Grifell-Tatjé E, Lovell CAK (2015) Productivity accounting: the economics of business performance. Cambridge University Press, New York
Kuosmanen T, Sipiläinen T (2009) Exact decomposition of the fisher ideal total factor productivity index. J Prod Anal 31:137–150
Mizobuchi H (2016) A superlative index number formula for the Hicks-Moorsteen productivity index. http://www.uq.edu.au/economics/cepa/docs/WP/WP03/2016.pdf
O'Donnell CJ (2012) An aggregate quantity framework for measuring and decomposing productivity change. J Prod Anal 38:255–272
Ray SC, Mukherjee K (1996) Decomposition of the fisher ideal index of productivity: a nonparametric dual analysis of U.S. airlines data. Econ J 106:1659–1678

Chapter 6
Productivity Interpretations of the Farrell Efficiency Measures and the Malmquist Index and Its Decomposition

Finn R. Førsund

Abstract The ratio definition of efficiency has the form of a productivity measure. But the weights are endogenous variables and they do not function well as weights in a productivity index proper. It is shown that extended Farrell measures of efficiency can all be given an interpretation as productivity measures as observed productivity relative to productivity at the various projection points on the frontier. The Malmquist productivity index is the efficiency score for a unit in a period relative to the efficiency score in a previous period, thus based on a maximal common expansion factor for outputs or common contraction factor for inputs not involving any individual weighting of outputs or inputs, as is the case if a Törnqvist or ideal Fisher index is used. The multiplicative decomposition of the Malmquist productivity index into an efficiency part and a frontier shift part should not be taken to imply causality. The role of cone benchmark envelopments both for calculating Malmquist indices of productivity change and for decomposing the indices into an efficiency change term and a frontier shift term is underlined, and connected to the index property of proportionality and circularity, adding the use of a fixed benchmark envelopment. The extended decomposition of the efficiency component by making use of scale efficiency is criticised.

Keywords Farrell efficiency measures · Technically optimal scale · Malmquist productivity index · Decomposition of the Malmquist productivity index

JEL Classification C18 · C43 · C61 · D24

F.R. Førsund (✉)
Department of Economics, University of Oslo, Oslo, Norway
e-mail: finn.forsund@econ.uio.no

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_6

6.1 Introduction

Measuring productive efficiency has been developing the last decades to become an important research strand within the fields of economics, management science and operations research. Two seminal contributions are Farrell (1957) and Charnes et al. (1978). Although the latter paper adopts the efficiency definition of the former the approaches for calculation the measure differ in the two papers. Farrell started out defining a frontier production function as the relevant comparison for measuring productive efficiency for observations of production units and introduced radial measures for the case of constant returns to scale. Charnes et al. (1978), formulating the optimisation problem for estimating the efficiency measure, set up a ratio of weighted outputs on weighted inputs. This approach brought the concept of productivity into the efficiency story. However, although the ratio formally looks like a productivity measure it is not set up to represent a productivity index proper, but to estimate efficiency, i.e. to compare the "productivity" of an observation with the productivity of a benchmark on the best practice frontier using weights that are endogenous. Using these weights, the weighted sum of inputs or outputs will be restricted to 1 (depending on estimating an output- or input-oriented efficiency score), and one or more weight may be zero contrary to what one would want constructing a productivity index proper.

A purpose of the chapter is to elaborate upon the productivity interpretation for the generalised Farrell efficiency measures covering the case of variable returns to scale. We then have technical efficiency measures, scale efficiency measures and a technical measure of productivity, the last two types of measures building upon the old concept of technically optimal scale in production theory. We will also have a closer look at the Malmquist productivity index because it is defined as the ratio of Farrell technical efficiency measures for a unit for two different time periods. A contribution of the chapter is to introduce some relevant concepts to an audience oriented toward DEA.

The chapter is organised as follows. The Charnes et al. (1978) ratio measure and five Farrell efficiency measures are defined in Sect. 6.2[1] and the productivity interpretations of the latter measures discussed for the case of a single output and input, and then generalised to multiple outputs and inputs. The importance of (local) constant returns to scale for productivity measurement is brought out using the elasticity of scale. In Sect. 6.3 the Malmquist index proposed in Caves et al. (1982) is introduced and some basic properties of the index and their consequences for choice of efficiency measures are discussed. The decomposition of productivity change into efficiency change and frontier shift introduced in Nishimizu and Page (1982) is discussed and compared with the decomposition done in Färe et al. (1992, 1994a, c). Section 6.4 offers some conclusions.

[1]Section 6.2 is based on Førsund (2015), Sect. 4.

6.2 Productivity Interpretations of the Farrell Efficiency Measures

6.2.1 The Ratio Definition of the Efficiency Measure

Charnes et al. (1978) relate the ratio idea for defining an efficiency measure to how efficiency is defined in engineering as "the ratio of the actual amount of heat liberated in a given device to the maximum amount that could be liberated by the fuel" (Charnes et al. 1978, p. 430). The optimisation problem set up for deriving the efficiency measure in the case of constant returns to scale (CRS) for a dataset, from a specific time period, is:

$$\begin{aligned} &\text{Max } h_{j_0} = \frac{\sum_{r=1}^{s} u_{rj_0} y_{rj_0}}{\sum_{i=1}^{m} v_{ij_0} x_{ij_0}} \quad \text{subject to} \\ &\frac{\sum_{r=1}^{s} u_{rj_0} y_{rj}}{\sum_{i=1}^{m} v_{ij_0} x_{ij}} \leq 1\,, j = 1\ldots, j_0, \ldots n,\ u_{rj_0},\ v_{ij_0} \geq 0\ \forall r, i \end{aligned} \tag{6.1}$$

Here h_{j0} is the efficiency measure for unit j_0, y_{j0} and x_{j0} are the output and input vectors, respectively, with s outputs and m inputs, number of units is n, and u_{rj0}, v_{ij0} are the weights for unit j_0 associated with outputs and inputs, respectively. These weights are endogenous variables and will be determined in the optimal solution. The constraints on the ratios in the optimisation problem (6.1) require the "productivity" of all units to be equal to or less than 1 using the weights for unit j_0, i.e. the productivity of fully efficient units is normalised to 1. Moreover, the weighted sum of inputs (input orientation) or outputs (output orientation) for the unit j_0 under investigation is normalised to 1 when the fractional programming problem (6.1) is converted to a linear programming problem as shown by Charnes et al. (1978), thus providing a link to the Farrell approach.[2]

6.2.2 The Farrell Suite of Efficiency Measures

Farrell (1957) defined two technical measures of efficiency, the input-oriented measure based on scaling inputs of inefficient units down with a common scalar, projecting the point radially to the frontier keeping observed output constant, and the output-oriented measure scaling outputs of inefficient units up with a common scalar, projecting the point radially to the frontier keeping observed inputs constant. The measures were defined for a frontier function exhibiting constant returns to

[2]Farrell and Fieldhouse (1962) were the first to solve the problem of calculating their efficiency measure by using linear programming.

scale.[3] However, he also discussed variable returns to scale and studied this further in Farrell and Fieldhouse (1962), without explicitly introducing measures reflecting scale properties. This was done in Førsund and Hjalmarsson (1974, 1979), developing a family of five efficiency measures. The latter paper illustrated the measures using a smooth variable returns to scale frontier production function exhibiting an *S*-shaped graph as typical for neoclassical production functions obeying the *Regular Ultra Passum Law*[4] of Frisch (1965).[5] However, the efficiency measures are valid for other types of frontier functions as long as a basic requirement of the variation of the elasticity of scale is fulfilled. In this paper the focus will be on a non-parametric piecewise linear frontier function; the generic DEA model exhibiting variable returns to scale (VRS) having a convex production possibility set, and exhibiting the other properties introduced in Banker et al. (1984).[6]

The family of the five Farrell efficiency measures is illustrated in Fig. 6.1 in the case of the frontier within a non-parametric framework being a piecewise linear convex function (Førsund 1992). The point of departure is the observation $P^0 = (y^0, x^0)$ that is inefficient with respect to the VRS frontier. The reference point on the frontier for the input-oriented measure E_1 with respect to the VRS frontier is $P_1^{VRS} = (y^0, x_1^{VRS})$, and the reference point on the frontier for the output-oriented measure E_2 with respect to the VRS frontier is $P_2^{VRS} = (y_2^{VRS}, x^0)$. A second envelopment is indicated by the ray from the origin being tangent to the point P^{Tops}. (I will return to the interpretation of this point below.) This frontier exhibits constant returns to scale (CRS). The reference points on the frontier are $P_1^{CRS} = (y^0, x_1^{CRS})$ and $P_2^{CRS} = (y_2^{CRS}, x^0)$. The dotted factor ray from the origin to the observation gives the productivity of the observation, and the dotted factor ray from the origin to a reference point on the VRS frontier gives the productivity of this reference point. As is easily seen from Fig. 6.1 the productivity at the CRS envelopment is the maximal productivity obtained on the VRS frontier. Comparing the observation with the reference point $P^{Tops} = (y^T, x^T)$ therefore gives the relative productivity of an observation to the maximal productivity on the VRS frontier. Continuing Farrell's numbering of measures a measure E_3 is introduced covering this measurement and is therefore termed the measure of *technical productivity*.[7]

[3]Farrell (1957) points out that the two measures in the case of constant returns to sale are equal.

[4]The Regular Ultra Passum Law requires that the scale elasticity decreases monotonically from values greater than one, through the value one to lower values when moving along a rising curve in the input space.

[5]This may be the reason for this way of presenting the family of efficiency measures being rather unknown in the DEA literature.

[6]In the VRS DEA specification the scale elasticity has a monotonically decreasing value in the range of increasing returns to scale, but has a more peculiar development in the range of decreasing returns to scale as shown in Førsund et al. (2009). However, there may be a unique face where the scale elasticity is equal to 1 along a rising curve, or else define a vertex point as having constant returns to scale when the left-hand elasticity at the point is less than one and the right-hand elasticity is greater than one.

[7]In Førsund and Hjalmarsson (1979), introducing this measure, it was called the gross scale efficiency.

The two remaining efficiency measures E_4 and E_5 introduced in Førsund and Hjalmarsson (1979) are the scale efficiency measures[8] comparing the productivity of the reference points P_1^{VRS} and P_2^{VRS}, respectively, with the point P^{Tops} of maximal productivity on the frontier.

6.2.3 *Productivity Interpretations in the Case of a Single Output and Input*

All Farrell measures of efficiency can be given an interpretation of relative productivity; the productivity of the observation relative to specific points on the VRS frontier marked in Fig. 6.1. Before showing the relative productivity interpretation in the case of a single output and a single input (Berg et al. 1992) in a general setting, let us state the definitions of the Farrell input-and output-oriented technical efficiency measures, starting with the general definition of the production possibility set $T = \{(y, x) : y \geq 0$ can be produced by $x \geq 0\}$ (y and x are vectors). By assumption let the set T exhibit variable returns to scale (VRS) of its frontier (the efficient boundary of T). The input-and output-oriented efficiency measures can be defined as[9]

$$\begin{aligned} E_1(y,x) &= \min_{\mu}\{\mu : (\mu x, y) \in T\} \\ E_2(y,x) &= \min_{\lambda}\{\lambda : (x, y/\lambda) \in T\}. \end{aligned} \tag{6.2}$$

The relative productivity interpretation can be shown in the case of a single output and input in the following way, starting with the input-oriented efficiency measure using the points P^0 and P_1^{VRS} in Fig. 6.1:

$$\frac{y^0/x^0}{y^0/x_1^{VRS}} = \frac{y^0/x^0}{y^0/E_1x^0} = E_1 \tag{6.3}$$

The same productivity interpretation holds for the output-oriented efficiency measure using points P^0 and P_2^{VRS} in Fig. 6.1:

$$\frac{y^0/x^0}{y_2^{VRS}/x^0} = \frac{y^0/x^0}{(y^0/E_2)/x^0} = E_2 \tag{6.4}$$

In the input-oriented case we adjust the observed input quantity so that the projection of the observation is on the frontier, and in the output-oriented case we

[8] In Førsund and Hjalmarsson (1979) these measures were called measures of pure scale efficiency.

[9] The Farrell efficiency measure functions correspond to the concept of distance functions introduced in Shephard (1970). Shephard's input distance function is the inverse of Farrell's input-oriented efficiency measure, and Shephard's output distance function is identical to Farrell's output-oriented efficiency measure.

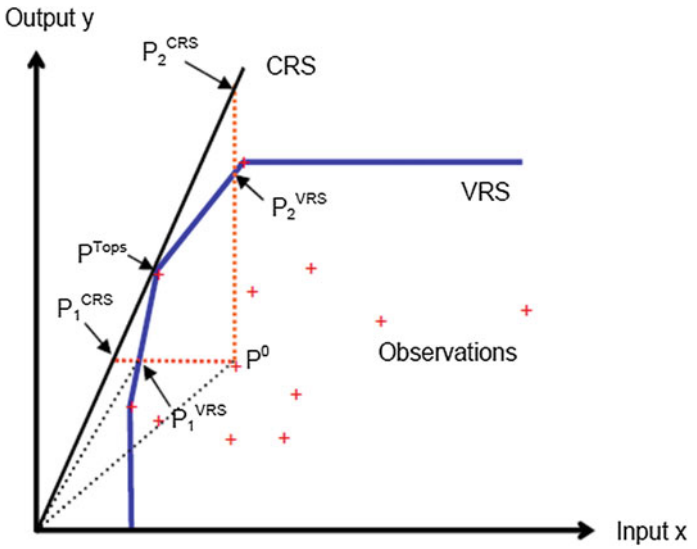


Fig. 6.1 The Farrell efficiency measures applied to a piecewise linear VRS frontier

adjust the observed output, using the symbols for adjusted input and output introduced above.

For the three remaining measures we will make a crucial use of the CRS envelopment in order to calculate the measures. The notation E_1^{CRS} and E_2^{CRS}, making explicit reference to the CRS envelopment as the frontier, together with $\mathrm{P}^{\mathrm{Tops}} = (y^{\mathrm{T}}, x^{\mathrm{T}})$, will be used. The measure of technical productivity is

$$\begin{aligned} \frac{y^0/x^0}{y^T/x^T} &= \frac{y^0/x^0}{y^0/E_1^{CRS}x^0} = E_1^{CRS} = E_3 \\ \frac{y^0/x^0}{y^T/x^T} &= \frac{y^0/x^0}{(y^0/E_2^{CRS})/x^0} = E_2^{CRS} = E_3 \Rightarrow E_3 = E_1^{CRS} = E_2^{CRS} \end{aligned} \tag{6.5}$$

The first expression in each of the two lines of the equations is the definition of the measure of technical productivity using the productivity at the point $\mathrm{P}^{\mathrm{Tops}}$ as a reference. The second expressions, input-orientation or output-orientation, respectively, show the most convenient way of calculating the productivity measure. The outputs and inputs differ between the observation P^0 and the $\mathrm{P}^{\mathrm{Tops}}$ points. But using the CRS envelopment the maximal productivity for the VRS technology is the same along the entire ray from the origin going through the point $\mathrm{P}^{\mathrm{Tops}}$. The productivity measure E_3 is equal to both the input-oriented measure and the output-oriented measure using the CRS envelopment as the frontier. It is easy to see geometrically that in the case of using the CRS envelopment the two orientated efficiency measures must be identical, as pointed out by Farrell (1957).

Measures for scale efficiency are also defined using a relative productivity comparison. The input-oriented scale efficiency E_4 (keeping output fixed) and the output-oriented scale efficiency E_5 (keeping input fixed) are:

$$\begin{aligned}\frac{y^0/x_1^{VRS}}{y^T/x^T} &= \frac{y^0/E_1x^0}{y^0/E_1^{CRS}x^0} = \frac{E_1^{CRS}}{E_1} = \frac{E_3}{E_1} = E_4 \\ \frac{y_1^{VRS}/x^0}{y^T/x^T} &= \frac{(y^0/E_2)/x^0}{y^0/E_2^{CRS}x^0} = \frac{E_2^{CRS}}{E_2} = \frac{E_3}{E_2} = E_5\end{aligned} \tag{6.6}$$

The relative productivity comparison for input-oriented scale efficiency in Fig. 6.1 is between the observed output on the efficiency-corrected input on the VRS frontier and the maximal productivity at the P^{Tops}-point (y^T, x^T). For output-oriented scale efficiency we have an analogous construction. The calculations of the scale efficiency measures can either be based on the ratios between the efficiency scores for input-oriented efficiency relative to the VRS frontier and the CRS envelopment, or expressed as deflating the technical productivity measure with the relevant efficiency measures relative to the VRS frontier. Notice that there is only a single technical efficiency measure for a CRS technology; we have $E_1 = E_2 = E_3$ and $E_4 = E_5 = 1$.

6.2.4 The Concepts of Elasticity of Scale and Technically Optimal Scale

Before generalising the relative productivity interpretation to multiple outputs and inputs we need to introduce the concept of elasticity of scale. The definition of a local scale elasticity for a frontier production function is the same whether it is of the neoclassical differential type $F(y, x) = 0$ or if the production possibility set has a faceted envelopment border like in the DEA case. We are looking at the maximal uniform proportional expansion β of outputs for a given uniform proportional expansion α of inputs, i.e. looking at $F(\beta y, \alpha x) = 0$. The local scale elasticity is defined as the derivative of the output expansion factor w.r.t. the input expansion factor on the average value of the ratio of the output factor on the input factor[10]:

$$\varepsilon(x, y) = \frac{\partial\beta(x, y, \alpha)}{\partial\alpha}\frac{\alpha}{\beta} = \frac{\partial\beta(\alpha, x, y)}{\partial\alpha}\Big|_{\alpha=\beta=1} \tag{6.7}$$

The scale elasticity is evaluated, without loss of generality, for $\alpha = \beta = 1$. In the DEA case with non-differentiable points (vertex points or points on edges) the expression above is substituted with the right-hand derivative or the left-hand derivative, respectively, at such points (Krivonozhko et al. 2004; Førsund and

[10]See Hanoch (1970), Panzar and Willig (1977), Starrett (1977).

Hjalmarsson 2004b; Førsund et al. 2007; Podinovski et al. 2009; Podinovski and Førsund 2010).

Returns to scale is defined by the value of the scale elasticity; increasing returns to scale is defined as $\varepsilon > 1$, constant returns to scale as $\varepsilon = 1$ and decreasing returns to scale as $\varepsilon < 1$.

For a production function with variable returns to scale there is a connection between the input- and output-oriented measures via the scale elasticity. Following Førsund and Hjalmarsson (1979) in the case of a frontier function for a single output and multiple inputs we have

$$E_2 = E_1^{\bar{\varepsilon}} \Rightarrow E_1 \underset{<}{\overset{>}{}} E_2 \text{ for } \bar{\varepsilon} \underset{<}{\overset{>}{}} 1 \tag{6.8}$$

where the variable $\bar{\varepsilon}$ is the average elasticity of scale along the frontier function from the evaluation point for the input-saving measure to the output-increasing measure. In Førsund (1996) this result was generalised for multiple outputs and inputs in the case of a differentiable transformation relation $F(y,x) = 0$ as the frontier function, using the *Beam [Ray] variation equations* of Frisch (1965). This result holds for points of evaluation being projection points in the relative interior of faces. The path between the points will be continuous although not differentiable at vertex point or points located at edges.

We must distinguish between scale elasticity and scale efficiency (Førsund 1996). Formalising the illustration in Fig. 6.1 the reference for the latter is the concept of *technically optimal scale* of a frontier function (Frisch 1965). The set of points $TOPS^T$ having maximal productivities for the (efficient) border of the production possibility set $T = \{(y,x) : y \geq 0 \text{ can be produced by } x \geq 0\}$ with the frontier exhibiting VRS, can be defined as (Førsund and Hjalmarsson 2004a)[11]

$$TOPS^T = \{(y,x) : \varepsilon(y, x) = 1, (y, x) \in T\} \tag{6.9}$$

It must be assumed that such points exist and that for outward movements in the input space the scale elasticity cannot reach the value of 1 more than once for a smooth neoclassical frontier. However, it can in the DEA case be equal to 1 for points on the same face (see footnote 6). The point (y^T, x^T) used above is now replaced by vectors y^T and x^T belonging to the set $TOPS^T$. From production theory we know that in general a point having maximal productivity must have a scale elasticity of 1. In a long-run competitive equilibrium the production units are assumed to realise the technically optimal scale with the scale elasticity of 1 implying zero profit.

[11]The concept of the M-locus in the case of multi-output was introduced in Baumol et al. (1982, pp. 58–59). In Førsund and Hjalmarsson (2004a) the M locus is defined and estimated within a DEA model using the TOPS set.

6.2.5 The Productivity Interpretation of the Efficiency Measures in the General Case

The interpretation of the five Farrell measures as measures of relative productivity can straightforwardly be generalised to multiple outputs and inputs. Introducing general aggregation functions[12] $Y = g_y\ (y_1,\ y_2,\ldots,y_M)$ and $X = g_x\ (x_1,\ x_2,\ldots,\ x_N)$ where Y and X are the scalars of aggregate quantities and $y_1, y_2,\ldots$ and $x_1, x_2,\ldots$ etc., are elements of the respective vectors y and x for outputs and inputs. The non-negative aggregation functions are increasing in the arguments and linearly homogeneous in outputs and inputs, respectively (O'Donnell 2012). We have, starting with the definition of relative productivity in the input-oriented case for an observation vector (y^0, x^0):

$$\underbrace{\frac{g_y(y^0)/g_x(x^0)}{g_y(y_1^{VRS})/g_x(x_1^{VRS})}}_{\text{Relative productivity}} = \underbrace{\frac{g_y(y^0)/g_x(x^0)}{g_y(y^0)/g_x(E_1x^0)}}_{\text{Substituting for frontier input}} = \underbrace{\frac{g_y(y^0)/g_x(x^0)}{g_y(y^0)/E_1g_x(x^0)}}_{\text{Using homogeneity of index function}} = E_1 \tag{6.10}$$

In the first expression relative productivity is defined in the input-oriented case using the observed vectors y^0, x^0 and the vectors y_1^{VRS}, x_1^{VRS} for the projection onto the VRS frontier analogous to the point P_1^{VRS} in Fig. 6.1 in the two-dimensional case. In the second expression the vectors for y_1^{VRS} and x_1^{VRS} are inserted, keeping the observed output levels y^0 and contracting the observed input vector using the input-oriented efficiency E_1 to project the inputs x^0 to the VRS frontier. In the third expression the homogeneity property of the input index function is used.

In the case of output orientation of the efficiency measure E_2 we get in the multiple output—multiple input case following the procedure above[13]:

$$\frac{g_y(y^0)/g_x(x^0)}{g_y(y_2^{VRS})/g_x(x_2^{VRS})} = \frac{g_y(y^0)/g_x(x^0)}{g_y(y^0/E_2)/g_x(x^0)} = \frac{g_y(y^0)/g_x(x^0)}{(g_y(y^0)/E_2)/g_x(x^0)} = E_2 \tag{6.11}$$

Using the general aggregation functions $g_y(y)$, $g_x(x)$ the measure E_3 of technical productivity can be derived using input- or output-orientation:

[12]Following the classical axiomatic (test) approach there are a number of properties (at least 20) an index should fulfil (Diewert 1992), the ones most often mentioned are monotonicity, homogeneity, identity, commensurability and proportionality. "Satisfying these standard axioms limits the class of admissible input (output) quantity aggregator functions to non-negative functions that are non-decreasing and linearly homogeneous in inputs (outputs)" (O'Donnell 2012, p. 257). There is no time index on the functions here because our variables are from the same period.

[13]The productivity interpretation of the oriented efficiency measures E_1 and E_2 can also be found in O'Donnell (2012, p. 259) using distance functions.

$$\frac{g_y(y^0)/g_x(x^0)}{g_y(y^T)/g_x(x^T)} = \frac{g_y(y^0)/g_x(x^0)}{g_y(y^0)/g_x(E_1^{CRS}x^0)} = \frac{g_y(y^0)/g_x(x^0)}{g_y(y^0)/E_1^{CRS}g_x(x^0)} = E_1^{CRS} = E_3$$
$$\frac{g_y(y^0)/g_x(x^0)}{g_y(y^T)/g_x(x^T)} = \frac{g_y(y^0)/g_x(x^0)}{g_y(y^0/E_2^{CRS})/g_x(x^0)} = \frac{g_y(y^0)/g_x(x^0)}{(g_y(y^0)/E_2^{CRS})/g_x(x^0)} = E_2^{CRS} = E_3$$
$$\Rightarrow E_1^{CRS} = E_2^{CRS} = E_3 \tag{6.12}$$

We obtain the same relationship between the technical productivity measure and the oriented measures with the CRS envelopment as in the simple case illustrated in Fig. 6.1. Notice that the use of points on the CRS envelopment in (6.12) is just introduced in order to calculate the measure E_3, and is not the basic definition of the measure; the definition is the first expression on the left-hand side of the two first lines.

The case of multi-output and -input is done in the same way for the scale efficiency measures as for the other measures utilising the homogeneity properties of the aggregation functions:

$$\begin{aligned}\frac{g_y(y^0)/g_x(x_1^{VRS})}{g_y(y^T)/g_x(x^T)} &= \frac{g_y(y^0)/g_x(E_1x^0)}{g_y(y^0)/E_1^{CRS}g_x(x^0)} = \frac{g_y(y^0)/E_1g_x(x^0)}{g_y(y^0)/E_1^{CRS}g_x(x^0)} \\ &= \frac{E_1^{CRS}}{E_1} = \frac{E_3}{E_1} = E_4 \\ \frac{g_y(y_2^{VRS})/g_x(x^0)}{g_y(y^T)/g_x(x^T)} &= \frac{g_y(y^0/E_2)/g_x(x^0)}{(g_y(y^0)/E_2^{CRS})/g_x(x^0)} = \frac{(g_y(y^0)/E_2)/g_x(x^0)}{(g_y(y^0)/E_2^{CRS})/g_x(x^0)} \\ &= \frac{E_2^{CRS}}{E_2} = \frac{E_3}{E_2} = E_5\end{aligned} \tag{6.13}$$

Again, we obtain the same relationship between the technical productivity measure and the oriented measures defining scale efficiency as in the simple case illustrated in Fig. 6.1. The calculations of the scale efficiency measures can either be based on the ratios between the efficiency scores for input-oriented efficiency relative to the VRS frontier and the CRS envelopment or expressed as deflating the technical productivity measure with the relevant efficiency measures relative to the VRS frontier.

6.3 The Malmquist Productivity Index

The point of departure is that we have observations of a set of the same units over time. The general construction of a total factor productivity index is to have an index for the volume of outputs over a volume index of inputs. A classical problem is to construct appropriate indices aggregating outputs and inputs, respectively.

The special feature of the Malmquist productivity index is that, without having any price data, volume indices can be established based on using efficiency scores for observations relative to estimated frontier production functions representing best practice. Caves et al. (1982) introduced the bilateral Malmquist productivity index developed for discrete time based on the ratio of distance functions (or Farrell efficiency functions that is the term used in this chapter) measured for two observations of the same unit at different time periods utilising efficiency scores only. Färe et al. (1994a, c) showed how to estimate the index in the case of specifying the possibility set as a convex polyhedral set and estimating the border of the set and efficiency scores using linear programming. The popularity soon followed. Caves et al. (1982) have 938 citations and Färe et al. (1994c) 929 in the Web of Social Science (per April 4, 2016).

However, the Malmquist productivity index was oriented, building on either an output-oriented efficiency score or an input-oriented one. A Malmquist index more of the traditional non-oriented type based on an index of output change over an index of input changes for two periods was introduced by Bjurek (1996), inspired by Diewert (1992) mentioning some ideas of Moorsteen and Hicks, hence the name Moorsteen-Hicks index adopted later, although Bjurek used the more functional name of a Malmquist total factor productivity index.[14]

However, a purpose with the present study is to look deeper into the decompositions of the original Caves et al. (1982) Malmquist productivity index, completely dominating in number of applications.[15]

6.3.1 *The Interpretation of the Malmquist Productivity Change Index*

The Caves et al. Malmquist oriented indices are utilising Farrell technical efficiency scores. The index for a unit i observed for two different time periods u and v, relative to the same border of the production possibility set indexed by b, representing one of the years, is:

$$M_{ij}^{b}(u,v)=\frac{E_j^b(x_{iv},y_{iv})}{E_j^b(x_{iu},y_{iu})},\ j=1,2,\quad i=1,\ldots,N,\quad u,v=1,\ldots,T,\quad u<v,\quad b=u,v \tag{6.14a}$$

The benchmark technology indexed by b is in many applications either the technology for period u or v, and changing over time according to the technology

[14]A thorough evaluation of the advantages of this type of a Malmquist productivity index is found in Lovell (2003), and it is also mentioned as the most satisfactory Malmquist type of productivity index in O'Donnell (2012), being what he called multiplicatively complete.

[15]Lovell (2003) decomposes also the Malmquist total factor productivity index multiplicatively into five terms. However, we will not investigate this issue here.

chosen as the base for the two periods involved. It is also usual to take a geometric mean of the results using technologies for both year u and v, following the seminal paper Färe et al. (1994a) on how to estimate the Malmquist productivity index.[16] The reason usually given in the literature is simply that either the technology from u or from v may be used as benchmark, and it is arbitrary which one to use, so the most reasonable is to take the geometric mean. As stated in Balk (1998, p. 59): "Since we have no preferences for either the geometric average of these index numbers will be used". Fare et al. (1994c, p.70) stated the reason as "In order to avod choosing an arbitrary benchmark". When a geometric mean is taken technologies for the two periods are involved, and when time moves forward this implies that the technology for a specific period is involved in two productivity change calculations (except for the first and last year).[17]

However, the time periods may be seen to impose a natural choice of the first period as a base in accordance with a "Laspeyres" view of using period u technology to gauge the productivity change from u to v. If the efficiency score for period v is greater (smaller) than the efficiency score for period u using period u technology, then there has been a productivity improvement (deterioration) from period u to period v.

It is well known in the literature how to set up LP problems to estimate the distance (or efficiency) functions involved in (6.14a) so we do not find it necessary to do this here (see e.g. Fried et al. 2008).

The efficiency functions in (6.14a) show the maximal proportional expansion (outputs) or contraction (inputs), and the measures are called technical efficiency measures because prices are not involved. The Malmquist productivity index is then a technical productivity index. There is no aggregation of outputs and inputs involved. Productivity change is measured as the relative change in the common expansion (contraction) factor between two periods.[18]

The productivity results may be different from the results one would get using prices for aggregating outputs and inputs. Weighting with revenue and cost shares as in the Törnqvist index means that the (real) price structure will have an influence. In general it seems more functional to choose weights according to importance placed on variables. The weights appearing in (1) are technically the dual variables, i.e. the shadow prices on output and input constraints (solving the "envelopment problem" in a DEA linear programming model) and give the marginal impact on the efficiency scores of changes in the exogenous observations, and are thus not related to the relative importance in an economic sense. Moreover, these shadow prices changes from one solution to the next in a more or less unpredictable manner. Using

[16]However, no reason is given for this procedure other than claiming that this was done in Caves et al. (1982). But there the geometric mean appears when establishing the connection between the Malmquist index and an Törnquist index assuming the unit to be on the frontier, while the fundamental assumption in Färe et al. (1994a) is that units may be inefficient.

[17]This may explain the empirical result in Bjurek et al. (1998) that productivity developments more or less follow each other for different formulations of the Malmquist index.

[18]The weighted ratio appearing in (1) should not be interpreted as productivity; the weights are just a by-product of the solutions of the optimisation problems in (6.2).

the ratio form as in (6.1) as a productivity index for the development between two time periods means that the weights are different for the solution of the efficiency scores for each period (Førsund 1998).

Another source of difference is that one or more of the weights of outputs and inputs in Eq. (6.1) may be zero, thus excluding variables from explicit influence on the efficiency scores in (6.14a) in order to maximise (minimise) the scaling factors in Eq. (6.2).[19] This may bias the Malmquist index in both directions compared with a standard Törnqvist index where all variables have strictly positive weights.

Another feature of the Malmquist productivity index that may give different results than other indices is that the efficiency functions in (6.14a) are based on frontier functions. In the case of capital vintage effects a dynamic investment process takes place in order to improve the technology level of a firm, so a frontier based on the best vintage technology may give a too optimistic view of the potential for efficiency improvements in the short run (Førsund 2010). The estimation of the frontier using DEA will also be distorted if observations picked to represent best practice by the method may in fact not be best practice, but picked due to biased technical change, as shown in Belu (2015), assuming a single vintage for each unit.

Thus, there is a question about the usefulness of the information a Malmquist productivity index gives compared with indices using available price information. Public sector production activities not selling outputs in markets seem to be the most relevant type of activities for application of the Malmquist productivity index.

In Sect. 6.2 the general aggregator functions $g_y(.)$ and $g_x(.)$ for outputs and inputs was introduced. These functions may now be period-specific. However, because we do not know these or do not have data to estimate them, the Malmquist index will be estimated using non-parametric DEA models giving us the efficiency measures in the numerator and denominator in (6.14a) (Färe et al. 2008).

When applying the Malmquist productivity index attention should be paid to desirable properties. In the literature this is more often than not glossed over. I will therefore explain in more detail the choice of the specification. Productivity as measured by the Malmquist index (6.14a) may be influenced by changes in the scale of the operation, but two units having the same ratio of outputs to inputs should be viewed as equally productive, regardless of the scale of production (Grifell-Tatjé and Lovell 1995). Doubling all inputs and outputs, keeping input and output mixes constant, should not change productivity. Therefore the benchmark envelopment of data, if we want to measure total factor productivity (TFP), is one that is homogenous of degree 1 in the input and output vectors, and thus the linear-homogenous set that fits closest to the data. The homogenous envelopment is based on the concept of technically optimal scale termed TOPS in Sect. 6.2. As pointed out in that section the productivity is maximal at optimal scale where returns to scale is one, thus the CRS contemporary benchmark envelopments

[19]To the best of my knowledge the pattern of occurrence of zero weights in Malmquist productivity index estimations has never been reported in the literature.

(assuming that the contemporaneous frontiers are VRS) are natural references for productivity changes over time.

In Fig. 6.2 observations of the same unit for the two periods u and v are indicated by P_u and P_v. The two corresponding VRS frontiers are drawn showing an outward shift indicating technological progress. The TOPS point for period v is labelled P_v^{Tops}. Just as the productivity should be unchanged if the input and output vectors are proportionally scaled, a measure of productivity should double if outputs are doubled and inputs are kept constant, and increase by half if inputs double, but outputs are constant. The desirable homogeneity properties of a TFP index is therefore to be homogenous of degree 1 in outputs in the second period v and of degree (-1) in inputs of the second period, and homogenous of degree (-1) in outputs of the first period u and homogenous of degree 1 in inputs of the first period. Using CRS to envelope the data is thus one way of obtaining all the required homogeneity properties of a Malmquist productivity change index. Notice that in the illustration in Fig. 6.2 the relative technology gap between the CRS benchmark technologies (blue lines) for observations in period v and u are identical, thus making the use of geometric mean of the Malmquist index in (6.14a) superfluous.[20]

The frontier technology level "jumps" from period to period from the start of one period to the start of the consecutive one. Outputs are produced and inputs consumed during the periods. This set-up is of course somewhat artificial compared with the fact that real changes take place in continuous time. The dynamic problems of adapting new technology and phasing it in are neglected. This theme is discussed in e.g. the literature on the Porter hypothesis and environmental regulation (Porter and van der Linde (1995); Brännlund and Lundgren 2009).[21]

Another property of a productivity index held to be important (Samuelson and Swamy 1974) is the *circularity* of the index (Berg et al. 1992; Balk and Althin 1996) (see Gini (1931) for an interesting exposition). The implied transitivity of the index means that the productivity change between two non-adjacent periods can be found by multiplying all the pairwise productivity changes of adjacent periods between the two periods in question, thus making identification of periods with weak or strong productivity growth possible. We will transitivise the Malmquist index by using a single reference frontier enveloping the pooled data, as illustrated by the upper (red) ray CRS(s) in Fig. 6.2. In Tulkens and van den Eeckaut (1995) this type of frontier was termed the *intertemporal frontier*.[22] Notice that taking the

[20]Most illustrations of the Malmquist indices in studies using geometric means are in fact using CRS frontiers and single output and input. Considering multiple outputs and inputs distances between contemporaneous frontiers will be independent of where the measure is taken if inverse homotheticity is assumed in addition to CRS, i.e. if Hicks neutral technical change is assumed.

[21]In panel data models efficiency change has been specified (Cornwell et al. 1990) as having unit-specific efficiencies that varies over time, but this is a "mechanical" procedure without an economic explanation of efficiency change.

[22]In Pastor and Lovell (2005), missing out on this reference, it was called the global frontier.

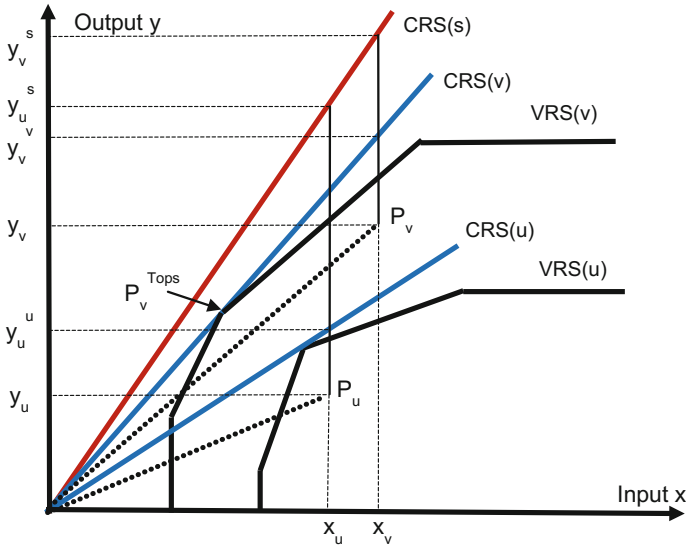


Fig. 6.2 The Malmquist productivity change index. Productivity change for a unit from period *u* to period *v* measured relative to the benchmark CRS(s) envelopment with maximal productivity of the pooled dataset

geometric mean of the Malmquist index (6.14a) for *u* and *v* used as benchmark envelopments is not compatible with circularity.

Using the same CRS reference envelopment for all units means that we have made sure that efficiency for all units and time periods refer to the same envelopment. The observations are either below the benchmark or on it in the case of the units from the pooled dataset spanning the envelopment. The pooled benchmark is identical to the *sequential frontier* of Tulkens and van den Eeckaut (1995) for the last period using sequentially accumulated data of all periods following the argument in Atkinson and Stiglitz (1969) that technology should not be forgotten.

Specifying CRS only is not sufficient to ensure that a specific data point occurring at different time periods get the same efficiency evaluation, because both input- and output isoquants may differ in shape over time if the technology is allowed to change over time as in Färe et al. (2008). The implication of having a time series of data is seldom discussed. Most illustrations and discussions seem to be focussed on two periods only. Hoverer, changing technologies successively as in (6.14a) implies that observations are measured against different frontiers over time. The question is the relevance for estimating productivity change of the information given by comparing relative numbers measured against different benchmarks.

Using a linear homogeneous envelopment implies that the orientation of the E function does not matter. The Malmquist index for a unit *i*, that should be used according to the properties outlined above is then:

$$M_i^s(u,v) = \frac{E^s(x_{iv}, y_{iv})}{E^s(x_{iu}, y_{iu})} = \frac{E_3^s(x_{iv}, y_{iv})}{E_3^s(x_{iu}, y_{iu})}, \quad i = 1, \ldots, J, \quad u, v = 1, \ldots, T, \; u < v \tag{6.14b}$$

where superscript s symbolises that all data are used for estimating the benchmark reference set. The productivity change is the change in the productivities of the observations relative to the benchmark maximal productivity, thus the E^s measures could formally be called E_3^s measures according to the terms introduced in Sect. 6.2, as done in the last expression in (6.14b). If all inputs are increased with a factor α and outputs with factor β from period u to period v ($\alpha x_{iu} = x_{iv}$ and $\beta y_{iu} = y_{iv}$) then we get from (6.14b): $M_i^s(u,v) = E_3^s(\alpha x_{iu}, \beta y_{iu})/E_3^s(x_{iu}, y_{iu}) = \beta/\alpha$; i.e. proportionality is obeyed due to the homogeneity properties of the efficiency score functions.

6.3.2 *The Decomposition of the Oriented Malmquist Productivity Index*

Nishimizu and Page (1982) were the first to introduce the decomposition of the productivity index into efficiency change and technical change in continuous time and then apply the decomposition in discrete time.[23] Färe et al. (1990, 1992, 1994a) adapted the decomposition to using a non-parametric frontier production function for estimating the efficiency scores. A quest for finding the sources of productivity change followed. I will return to some of these efforts after reviewing the decomposition of Nishimizu and Page (1982) that seems to be overlooked. They were aware of the problems with interpretation in the discrete case:

> Clearly, technological progress and technical efficiency change are not neatly separable either in theory or in practice. In our methodological approach […] we define technological progress as the movement of the best practice or frontier production over time. We then refer to all other productivity change as technical efficiency change. The distinction which we have adopted is therefore somewhat artificial, […]. (Nishimizu and Page (1982), pp. 932–933)

Their approach is set out in Fig. 6.3 (the original Fig. 1, p. 924). All variables are measured in logarithms, and the frontier functions are linear C–D functions with Hicks-neutral technical change from period 1 to period 2. Production is x and input z. The observation A has a production function with the same parameter as the frontiers g_1 and g_2, but with a different constant term. It is then the case that if unit A in period 1 had had the input of period 2, its production level would be at point B. From this point the frontier gap bc is added ending in point C', so $BC' = bc$.

[23] Nishimizu and Page (1982) were the first to refer to a working paper (Caves et al. 1981) that was published as Caves et al. (1982). However, they did not themselves use the term Malmquist productivity index.

Now, the observation in period 2 is found at C greater than C′. Nishimizu and Page then assume that the full potential frontier shift is realised in period 2, but in addition there is a positive efficiency change equal to *C′C*. So, measured in logarithms the productivity change is the sum of the efficiency gap C′C and the frontier gap *BC′* (=*bc*).

Figure 6.4 provides an explanation of their approach in the usual setting of quantities of variables in the simple case of single output and input and the frontiers being CRS. I will now show that the Nishimizu and Page decomposition is the same as the decomposition introduced in Färe et al. (1990, 1992, 1994a, c). A unit is observed at *b* in period 1 and at *f* in period 2. Using the frontier 1 as the benchmark technology instead of the pooled data for all years for simplicity of comparison the Malmquist productivity index (6.14b) for a unit *i* for change between period 1 and 2 and its decomposition are:

$$M_i^1(1,2) = \frac{E^1(y_i^2, x_i^2)}{E^1(y_i^1, x_i^1)} = \frac{E^2(y_i^2, x_i^2)}{E^1(y_i^1, x_i^1)} \cdot \frac{E^1(y_i^2, x_i^2)}{E^2(y_i^2, x_i^2)} = MC_i \cdot MF_i,$$
$$\frac{df/de}{ab/ac} = \frac{df/dg}{ab/ac} \cdot \frac{df/de}{df/dg},\ MF = \frac{df/de}{df/dg} = \frac{dg}{de} \tag{6.15}$$

The general definition of the Malmquist productivity-change index after the first equality sign is the ratio of the period efficiency measures against the same frontier technology, here for period 1. The expression after the second equality sign shows the multiplicative decomposition into a catching-up[24] measure *MC* and a frontier shift measure *MF*. The second line relates the observations *b* and *f* in Fig. 6.4 to the decomposition in the case of a single output and input. To obtain the correct homogeneity properties we have to use period frontiers that exhibit CRS. We are after information on sources for changes in the Malmquist productivity index, so even if the true contemporary frontier is VRS this does not mean that this frontier is the relevant one to use for the decomposition. I will return to this in the next subsection.

The *MF*-measure represents the relative gap between technologies and is thus the *potential* maximal contribution of new technology to productivity change, while the *MC*-measure is residually determined and may not represent the real efficiency contribution to productivity change, but illustrates the catching-up that is also influenced by the technology shift. It should be observed that the decomposition terms are *multiplied* to give the Malmquist index and not added.

Given that the only information we have about productivity change is the movement of an observation in input—output space, to distinguish between efficiency and technical change is rather difficult. The split into efficiency change and frontier shift that Nishimizu and Page proposed, is, concerning *MF*, based on assuming that the full productivity potential of the frontier shift is actually realised. If both observations had been on their respective frontiers it is obvious that the

[24]To the best of my knowledge this term was first used in Førsund (1993), and then in Fare et al. (1994c).

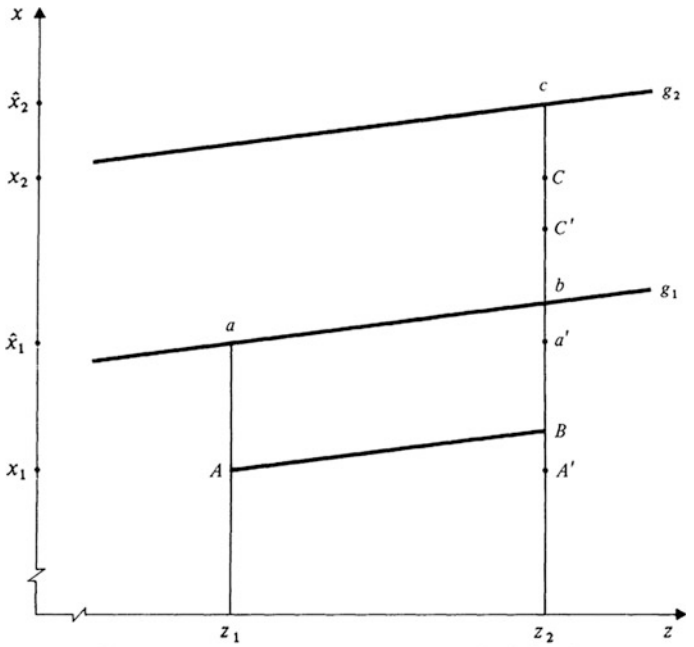


Fig. 1. Alternative interpretations of productivity change.

Fig. 6.3 The Nishimizu and Page (1982) decomposition. *Source* The Economic Journal

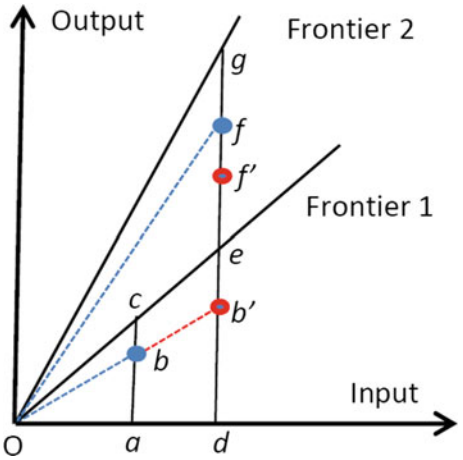


Fig. 6.4 The decomposition of the Malmquist index

Malmquist productivity change will reflect the frontier shift only. If both observations are inefficient with respect to their period frontiers then the efficiency contribution is measured by changing (expanding in Fig. 6.4) the input *Oa* in period

1 to that of *Od* in period 2, but using the actual production function in use in period 1 to predict the hypothetical output level at *f′*. However, I do not operate with any production function for an inefficient observation as Nishimizu and Page did (a CRS C–D function with the same form as the frontier functions), but I will equivalently assume that the efficiency level stays constant getting the inputs of period 2 in period 1. The unit then moves from point *b* to point *b′*. The problem is now to predict where observation *b′* in period 2 will be if the whole potential shift is realised as productivity change. Nishimizu and Page operated with logarithms of the variables and could more conveniently illustrate this, as shown in Fig. 6.3 above. In our Fig. 6.4 this means that the predicted output at point *f′* must obey $df'/db' = dg/de$, the latter being the relative frontier gap Then the same measure for efficiency "contribution" is actually obtained as in Nishimizu and Page, equal to the ratio of the two period efficiency measures. This decomposition is the same as the decomposition introduced in Färe et al. (1990, 1992, 1994a, c). This can be demonstrated in Fig. 6.4 by identifying the efficiency gap as df/df' and the frontier gap df'/db' building on Fig. 1 in Nishimizu and Page (Fig. 6.3 here), and using $df'/db' = dg/de$ and $db'/de = ab/ac$:

$$\frac{df}{df'} \cdot \frac{df'}{db'} = \frac{df/dg}{db'/de} \cdot \frac{dg}{de} = \frac{df/dg}{ab/ac} \cdot \frac{dg}{de} = \frac{df/de}{ab/ac} = M \tag{6.16}$$

However, note that the decomposition does not mean that there is a causation; we cannot distinguish between productivity change due to increase in efficiency and due to shift in technology using the general components in (6.15), as may seem to be believed in some of the empirical literature. The actual productivity change that we estimate using the Malmquist productivity index is from the observation in one period to an observation in another period (from *b* to *f* in Fig. 6.4). The causation is another question related to the dynamics of technical change and how this potential is utilised. As expressed in Nishimizu and Page (1982) after identifying technological progress as the change in the best practice production frontier:

> We then refer to all other productivity change – for example learning by doing, diffusion of new knowledge, improved managerial practice as well as short run adjustment to shocks external to the enterprise – as technical efficiency change. Nishimizu and Page (1982, p. 921)

Nishimizu and Page consider that dynamic factors influence efficiency change, but do not consider the same for realising the new technology.

We cannot decompose efficiency effects and frontier shift effects without making assumptions, according to Nishimizu and Page. Catching up seems to be the best descriptive term for the efficiency component. The decomposition can then be described as the relative potential contribution from technical change multiplied by an efficiency correction factor.

6.3.3 Circularity and Decomposition

Maintaining circularity for both components *MC* and *MF* in the decomposition implies that the technology shift term *MF* will be more complicated. Efficiency measures calculated relative to the benchmark frontier must be involved in the frontier shift measure. A decomposition of the index in Eq. (6.14b) that functions is:

$$M_i^s(u,v) = \underbrace{\frac{E^v(x_{iv},y_{iv})}{E^u(x_{iu},y_{iu})}}_{MC} \cdot \underbrace{\frac{E^s(x_{iv},y_{iv})/E^v(x_{iv},y_{iv})}{E^s(x_{iu},y_{iu})/E^u(x_{iu},y_{iu})}}_{MF},\ i=1,\ldots,N,\ u,v=1,\ldots,T,\ u<v \tag{6.17}$$

The *MF* measure of technology shift is calculated as a 'double' relative measure where both period efficiency measures are relative to the benchmark efficiency measures (Berg et al. 1992). It is easy to see that the decomposition reduces to the Malmquist index (6.14b) by cancelling terms. Notice that to do the decomposition we need benchmark envelopments for each period to be compared in addition to the fixed benchmark envelopment as seen in Fig. 6.2.

It can be illustrated in the case of one output and one input that the frontier shift component still measure the gap between the two benchmark technologies 1 and 2 in Figs. 6.2 and 6.4. Introducing the intertemporal benchmark *s* in Fig. 6.4 we can express the Malmquist index and its components in Fig. 6.5. The observations in period 1 and 2 are marked with blue circles at *b* and *h*. The relative frontier gap between frontier 1 and 2 measured using the observation for period 2 is *fk*/*fg*. We shall see if the decomposition in (6.17) gives the same measure using the notation in Fig. 6.5:

$$M = \frac{fh/fm}{ab/ae} = \underbrace{\frac{fh/fk}{ab/ac}}_{MC} \cdot \underbrace{\frac{(fh/fm)/(fh/fk)}{(ab/ae)/(ab/ac)}}_{MF} \tag{6.18}$$

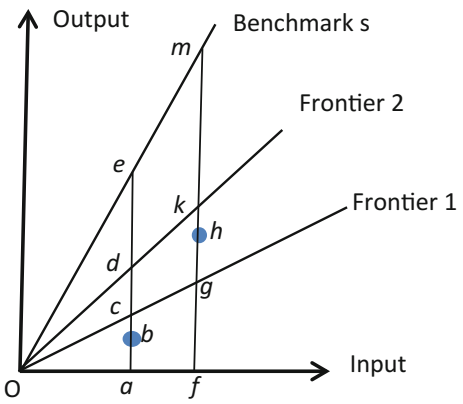


Fig. 6.5 The decomposition of the Malmquist index imposing circularity

The *MF* component can be developed as follows:

$$MF = \frac{(fh/fm)/(fh/fk)}{(ab/ae)/(ab/ac)} = \frac{fk/fm}{ac/ae} \tag{6.19}$$

The last expression is the gap between frontier 2 and benchmark s in the numerator and the gap between frontier 1 and the benchmark in the denominator, both expressed as the inverse of the definition of the gap as expressed in the last equation in (6.15). But using the property of like triangles we have $ac/ae = fg/fm$. The last expression in (6.19) can then be written:

$$\frac{fk/fm}{ac/ae} = \frac{fk/fm}{fg/fm} = \frac{fk}{fg} \tag{6.20}$$

This is the relative gap between frontier 2 and 1 using the input for period 2 as the base for calculating the gap.

However, note that in the general multi-output—multi-input case we cannot invoke the property of like triangles; the relative gaps depend on the input and output mixes.

6.3.4 Comments on Decompositions

In Färe et al. (1994b, c) the decomposition into catching up and frontier shift in Färe et al. 1990, 1992, 1994a)[25] was extended to a further decomposition of the efficiency change term into a scale efficiency term and a technical efficiency term, assuming the two contemporaneous frontiers to be VRS. This approach was criticised in Ray and Desli (1997) and a reply given in Färe et al. (1997). In his extensive review of decompositions of the Malmquist index Lovell (2003, p. 442) states: "I conclude that the Färe et al. (1994c) decomposition of the Malmquist productivity index is inadequate".

However, there are problems with the extended decompositions that are not observed by any of the papers above. The first comment is that decompositions are meant to identify sources of impacts on the total factor productivity index of observed movements of a unit in input-output space. It is then not necessarily the

[25]The history of the DEA-based Malmquist productivity index is presented in Färe et al. (1998), Grosskopf (2003) and Färe et al. (2008). The first working paper that established an estimation procedure based on DEA was published in 1989, was presented at a conference in Austin in the same year, and appeared as Färe et al. (1994a); a book chapter in a volume containing many of the conference presentations. The first journal publication appeared as Färe et al. (1990) with an application to electricity distribution. (However, this paper is not referred to in the 2003 and 2008 reviews and neither in Färe et al. (1992), although the methodological approach in the latter is the same).

case that one should use the actual contemporaneous technologies as a point of departure. A point that is under-communicated is the role of the benchmark envelopment. If we want the productivity change index to have the fundamental property of proportionality, then this envelopment have to exhibit constant returns to scale (CRS) even though the true contemporaneous technology is variable returns to scale. It follows most naturally that the decompositions should then also be based on envelopments exhibiting CRS. Thus, I support the choice in Färe et al. (1994c) of a cone benchmark envelopment. Ray and Desli (1997) do not seem to understand this choice, and in Balk (2001, p. 172) it is stated "it is not at all clear why technical change should be measured with respect to the cone technology" in spite of introducing proportionality in his Eq. (2).

Figure 6.2 illustrates the situation; the true contemporaneous technologies may be the variable returns to scale (VRS) functions for the two years, while the benchmark envelopment is represented by the cone CRS(s) based on the pooled data. Now, the catching-up term is the relative distance to the cone envelopments of the data from the two periods, while the frontier shift component is the "double relativity" format of (6.17) also involving distances from the observations to the benchmark envelopment of the pooled data.

There are many followers of the extended multiplicative decomposition in Färe et al. (1994b, c) of decomposing the catching-up term into what is called "pure" technical efficiency and scale efficiency. Pure efficiency is defined as the efficiency of the observation relative to the VRS frontier termed E_2 in Sect. 6.2. Using the terms there we have $E_3 = E_2 \cdot E_5$.[26] The complete decomposition of the change in the catching-up term, assuming a VRS technology for periods u and v and simplifying the notation, dropping writing the variables and unit index, is then

$$\frac{E_{3v}^{v}}{E_{3u}^{u}} = \frac{E_{2v}^{v}}{E_{2u}^{u}} \cdot \frac{E_{5v}^{v}}{E_{5u}^{u}} \tag{6.21}$$

However, it is difficult to see that this decomposition is helpful in interpreting the catch-up term. It is difficult to consider this term as a "driving factor". The E_2 terms are just there to satisfy the identity. The period VRS frontiers do not contribute to the understanding of the productivity changes based on CRS benchmark envelopments constructed by the analyst focusing on the development of the maximal productivity over time. The catch-up term is based on the change in the optimal scale (TOPS). Scale inefficiency has no role in our measure of productivity change. As remarked by Kuosmanen and Sipiläinen (2009, p. 140) "the distinction between the technical change and scale efficiency components is generally ambiguous and debatable." In Balk (2001) change in input mix is also identified as a separate factor, cf. O'Donnell (2012) also including change in output mix. However, these factors are not considered here.

[26] As a control, inserting the definition of E_5 we have for each period technology $E_3 = E_2 \cdot E_3/E_2 = E_3$.

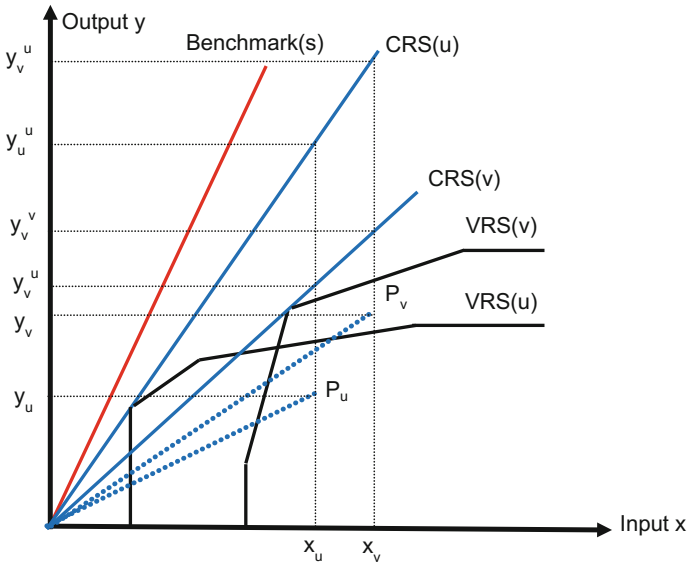


Fig. 6.6 Contemporaneous cones and VRS technologies

From a computational point of view the Malmquist index (6.14b) connects data points with a benchmark envelopment that serves our purpose of measuring productivity change. Pooling the data secures the highest possible degrees of freedom. Decomposition requires the estimation of two additional contemporaneous benchmark envelopments, and reduces the degrees of freedom in the estimation of these in addition to not giving us that much information of real sources behind productivity change.

We may face trouble also with basing the decomposition terms on cone envelopments if estimations of period functions are not properly restricted. An example is given in Fig. 6.6. The VRS envelopments are changed from those in Fig. 6.2 and are crossing each other,[27] and in such a way that the productivity of the optimal scale in period u is greater than in the later period v. We see clearly that the productivity growth measured by the Malmquist index (6.14b) shows growth, but that the frontier shift between periods u, v will show technical regress ($MF < 1$). However, the catching-up component then has to be greater than 1, and so much greater that growth is shown by the product of the terms. Looking at the VRS frontiers where the observations are located conveys that there has been a positive shift in the frontier from period u to period v, but this is the opposite of what the change in the period CRS benchmark tells us. One way to avoid this situation is to use sequential period envelopments. Then the CRS envelopment for period u may

[27]Crossing of technologies and crossing of isoquants as illustrated in Førsund (1993) will be difficult to interpret using geometric means of an index of the type in (6.14a).

be the same as for period v in Fig. 6.6 and productivity growth will be measured as due to efficiency improvement only.

6.4 Conclusions

Efficiency and productivity are two different concepts, but related through the fundamental definition of efficiency as being the relative relationship between the observed productivity of a unit and the maximal achievable productivity for the type of activity in question. Charnes et al. (1978) set up a different route to calculate the same efficiency measures introduced by Farrell (1957) by setting up a ratio form of productivity measures for estimating the efficiency scores, where the weights in the linear aggregation of outputs and inputs are estimated when maximising weighted outputs on weighted inputs subject to no productivity ratio using these weights for all units being greater than one (as a normalisation). However, this way of defining efficiency measures using expressions formally equal to productivity, is not as satisfactory for economists as the Farrell approach, introducing explicitly a frontier production function as a reference for efficiency measure definitions and calculations.

The original Farrell measures developed for constant returns to scale (CRS) has been extended to five efficiency measures for a frontier production function exhibiting variable returns to scale (VRS); input- and output technical efficiency, input- and output scale efficiency, and the technical productivity measure. The relationship between the two measures of technical efficiency involves the average scale elasticity value between the two frontier projection points along the frontier surface. The technical productivity measure and the two scale efficiency measures are developed based on the Frisch (1965) concept of technically optimal scale, predating the use of the concept most productive scale size in the DEA literature with almost 20 years.

It does not seem to be commonly recognised in the DEA literature that in the general case of multiple outputs and inputs the Farrell efficiency measures can all be given productivity interpretations in a more satisfactory way than the ratio form of Charnes et al. (1978). Using quite general theoretical aggregation functions for outputs and inputs with standard properties, it has been shown that all five Farrell efficiency measures can be given a productivity interpretation employing a proper definition of productivity. Each of the two technical efficiency measures and the technical productivity measure can be interpreted as the ratio of the productivity of an inefficient observation and the productivity of its projection point on the frontier, using the general aggregation equations. Of course, we have not estimated any productivity index as such, this remains unknown, but that was not the motivation of the exercise in the first place.

The Malmquist productivity index for bilateral comparisons, applied to discrete volume data and no prices, utilises Farrell efficiency measures directly. In order to have the required index property of proportionality it is sufficient to have as a

benchmark an envelopment that exhibits global constant returns to scale, although the underlying contemporaneous production frontiers may have variable returns to scale. One way of obtaining the proportionality properties is basing the benchmark envelopment on the technically optimal scale of the underlying frontiers. If circularity is wanted then this may be done by using cone envelopment for a single year, or pooling all data and using an intertemporal benchmark as is followed in this paper.

Fundamental drivers of productivity change are improvement in efficiency and technical change. The question is how to identify these drivers for a given dataset of outputs and inputs for units. The seminal contribution in Nishimizu and Page (1982) showed one way decomposing a productivity index into a component expressing efficiency change and a component showing the frontier shift impact on productivity that is shown to be the same type of decomposition as the one done for the Malmquist index of productivity change in Färe et al. (1994a). However, a warning of not attaching causality to the decomposition is in place. The decomposition is based on assuming that the full potential of productivity change due to new technology is actually realised, and then the efficiency component is determined residually, but neatly expressed as the relative catching-up to the last period frontier compared with the relative distance to the frontier in the previous period.

If a total factor productivity change index is wanted it is shown that a cone benchmark envelopment satisfy the proportionality test and furthermore using a fixed benchmark technology, for instance based on the pooled dataset as done in this chapter, will satisfy the circularity test. Furthermore, it is argued that cone benchmark envelopments should also be used for contemporaneous frontiers, thus criticising efforts to do further decompositions involving scale efficiencies based on assuming variable returns to scale period frontiers.

References

Atkinson AB, Stiglitz JE (1969) A new view of technological change. Econ J 79(315):573–578

Balk BM (1998) Industrial price, quantity, and productivity indices: the micro-economic theory and an application. Kluwer Academic Publishers, Boston-Dordrecht-London

Balk BM (2001) Scale efficiency and productivity change. J Prod Anal 15(3):159–183

Balk BM, Althin R (1996) A new, transitive productivity index index. J Prod Anal 7(1):19–27

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiency in data envelopment analysis. Manage Sci 30(9):1078–1092

Baumol WJ, Panzar JC, Willig RD (1982) Contestable markets and the theory of industry structure. Harcourt Brace Jovanovich, New York

Belu C (2015) Are distance measures effective at measuring efficiency? DEA meets the vintage model. J Prod Anal 43(3):237–248

Berg SA, Førsund FR, Jansen ES (1992) Malmquist indices of productivity growth during the deregulation of Norwegian banking, 1980-89. Scand J Econ 94(Supplement):S211–S228

Bjurek H (1996) The Malmquist total factor productivity index. Scand J Econ 98(2):303–313

Bjurek H, Førsund FR, Hjalmarsson L (1998) Malmquist productivity indexes: an empirical comparison. In Färe R, Grosskopf S, Russell RR (eds) Index numbers: essays in honour of Sten Malmquist. Kluwer Academic Publishers, Boston, Essay 5, pp 217–239

Brännlund R, Lundgren T (2009) Environmental policy without cost? A review of the Porter hypothesis. Int Rev Environ Resour Econ 3:75–117

Caves DW, Christensen LR, Diewert E (1981) A new approach to index number theory and the measurement of input, output, and productivity. Discussion Paper 8112, Social Systems Research Institute, University of Wisconsin

Caves DW, Christensen LR, Diewert E (1982) The economic theory of index numbers and the measurement of input, output, and productivity. Econometrica 50(6):1393–1414

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2(6):429–444

Cornwell C, Schmidt P, Sickles R (1990) Production frontiers with cross-sectional and time-series variation in efficiency levels. J Econometrics 46(1–2):185–200

Diewert WE (1992) Fisher ideal output, input and productivity indexes revisited. J Prod Anal 3 (3):211–248

Färe R, Grosskopf S, Yaisawarng S, Li SK, Wang Z (1990) Productivity growth in Illinois electric utilities. Resour Energy 12:383–398

Färe R, Grosskopf S, Lindgren B, Roos P (1992) Productivity changes in Swedish pharmacies 1980-1989: a non-parametric Malmquist approach. J Prod Anal 3(1/2):85–101

Färe R, Grosskopf S, Lovell CAK (1994b) Production frontiers. Cambridge University Press, Cambridge

Färe R, Grosskopf S, Lindgren B, Roos P (1994a) Productivity developments in Swedish hospitals: a Malmquist output index approach: In Charnes A, Cooper WW, Lewin AY, Seiford LM (eds) Data envelopment analysis: theory, methodology, and applications. Kluwer Academic Publishers, Boston/Dordrecht/London, Chapter 13, pp 253–272. (Also published in 1989 as *Discussion paper 89–3*, Department of Economics, Southern Illinois University, Carbondale)

Färe R, Grosskopf S, Norris M, Zhang Z (1994c) Productivity growth, technical progress and efficiency change in industrialized countries. Am Econ Rev 84(1):66–83

Färe R, Grosskopf S, Norris M (1997) Productivity growth, technical progress, and efficiency change in industrialized countries: reply. Am Econ Rev 87(5):1040–1044

Färe R, Grosskopf S, Roos P (1998) Malmquist productivity indexes: a survey of theory and practice. In Färe R, Grosskopf S, Russell RR (eds) Index numbers: essays in honour of Sten Malmquist. Kluwer Academic Publishers, Boston/London/Dordrecht, Essay 3, pp 127–190

Färe R, Grosskopf S, Margaritis D (2008) Efficiency and productivity: Malmquist and more. In Fried HO, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency and productivity growth. Oxford University Press, New York, Chapter 5, pp 522–622

Farrell MJ (1957) The measurement of productive efficiency. J Roy Stat Soc Ser A (General) 120 (3):253–281(290)

Farrell MJ, Fieldhouse M (1962) Estimating efficient production functions under increasing returns to scale. J Roy Stat Soc Ser A (General) 125(2):252–267

Førsund FR (1992) A comparison of parametric and non-parametric efficiency measures: the case of Norwegian ferries. J Prod Anal 3(1):25–43

Førsund FR (1993) Productivity growth in Norwegian ferries. In Fried HO, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency. Oxford University Press, Oxford, Chapter 14, pp 352–373

Førsund FR (1996) On the calculation of the scale elasticity in DEA models. J Prod Anal 7 (3):283–302

Førsund FR (1998) The rise and fall of slacks: comments on Quasi-Malmquist productivity indices. J Prod Anal 10(1):21–34

Førsund FR (2010) Dynamic efficiency measurement. Indian Econ Rev 45(2):125–159

Førsund FR (2015) Economic perspectives on DEA. Memorandum No 10/2015 from the Department of Economics, University of Oslo

Førsund FR, Hjalmarsson L (1974) On the measurement of productive efficiency. Swed J Econ 76 (2):141–154
Førsund FR, Hjalmarsson L (1979) Generalised Farrell measures of efficiency: an application to milk processing in Swedish dairy plants. Econ J 89(354):294–315
Førsund FR, Hjalmarsson L (2004a) Are all scales optimal in DEA? Theory and empirical evidence. J Prod Anal 21(1):25–48
Førsund FR, Hjalmarsson L (2004b) Calculating scale elasticity in DEA models. J Oper Res Soc 55(10):1023–1038. doi:10.1057/palgrave.jors.2601741
Førsund FR, Hjalmarsson L, Krivonozhko VE, Utkin OB (2007) Calculation of scale elasticities in DEA models: direct and indirect approaches. J Prod Anal 28(1):45–56
Førsund FR, Kittelsen SAC, Krivonozhko VE (2009) Farrell revisited–visualizing properties of DEA production frontiers. J Oper Res Soc 60(11):1535–1545
Fried HO, Lovell CAK, Schmidt SS (eds) (2008) The measurement of productive efficiency and productivity growth. Oxford University Press, New York
Frisch R (1965) Theory of production. D. Reidel, Dordrecht
Gini C (1931) On the circular test of index numbers. Metron 9(2):3–24
Grifell-Tatjé E, Lovell CAK (1995) A note on the Malmquist productivity index. Econ Lett 47 (2):169–175
Grosskopf S (2003) Some remarks on productivity and its decompositions. J Prod Anal 20(3):459–474
Hanoch G (1970) Homotheticity in joint production. J Econ Theor 2(4):423–426
Krivonozhko V, Volodin AV, Sablin IA, Patrin M (2004) Constructions of economic functions and calculation of marginal rates in DEA using parametric optimization methods. J Oper Res Soc 55(10):1049–1058
Kuosmanen T, Sipiläinen T (2009) Exact decomposition of the Fisher ideal total factor productivity index. J Prod Anal 31(3):137–150
Lovell CAK (2003) The decomposition of Malmquist productivity indexes. J Prod Anal 20 (3):437–458
Nishimizu M, Page JM (1982) Total factor productivity growth, technological progress and technical efficiency change: dimensions of productivity change in Yugoslavia 1965-78. Econ J 92(368):920–936
O'Donnell CJ (2012) An aggregate quantity framework for measuring and decomposing productivity change. J Prod Anal 38(3):255–272. doi:10.1007/s11123-012-0275-1
Panzar JC, Willig RD (1977) Economies of scale in multi-output production. Quart J Econ 91 (3):481–493
Pastor JT, Lovell CAK (2005) A global Malmquist productivity index. Econ Lett 88(2):266–271
Podinovski VV, Førsund FR (2010) Differential characteristics of efficient frontiers in data envelopment analysis. Oper Res 58(6):1743–1754
Podinovski VV, Førsund FR, Krivonozhko VE (2009) A simple derivation of scale elasticity in data envelopment analysis. Eur J Oper Res 197(1):149–153
Porter ME, van der Linde C (1995) Toward a new conception of the environment-competitiveness relationship. J Econ Perspect 9(4):97–118
Ray SC, Desli E (1997) Productivity growth, technical progress and efficiency change in industrialized countries: comment. Am Econ Rev 87(5):1033–1039
Samuelson PA, Swamy S (1974) Invariant economic index numbers and canonical duality: survey and synthesis. Am Econ Rev 64(4):566–593
Shephard RW (1970) Theory of cost and production functions (first edition 1953). Princeton University Press, New Jersey
Starrett DA (1977) Measuring returns to scale in the aggregate, and the scale effect of public goods. Econometrica 45(6):1439–1455
Tulkens H, van den Eeckaut P (1995) Non-parametric efficiency, progress, and regress measures for panel data: methodological aspects. Eur J Oper Res 80(3):474–499

Chapter 7
On the Use of DEA Models with Weight Restrictions for Benchmarking and Target Setting

Nuria Ramón, José L. Ruiz and Inmaculada Sirvent

Abstract This chapter discusses the use of DEA models with weight restrictions for purposes of benchmarking and target setting. Weight restrictions have been used in the literature to incorporate into the analysis both value judgments (managerial preferences or prior views about the relative worth of inputs and outputs) and technological judgments (assumptions on production trade-offs). An important consideration in the use of restricted models for the benchmarking is that they may provide targets that are outside the production possibility set (PPS). Such difficulties are overcome if weight restrictions represent production trade-offs, because in those cases restricted models lead to a meaningful expansion of the production technology. However, if weight restrictions are only used as a way of incorporating preferences or value judgments, then there could be no reason to consider the targets derived from those models as attainable. Despite the classical restricted DEA formulations may yield targets within the PPS, it is claimed here that an approach based on a more appropriate selection of benchmarks would be desirable. We develop some restricted models which provide the closest targets within the PPS that are Pareto-efficient. Thus, if weight restrictions represent value judgments, the proposed approach allows us to identify best practices which show the easiest way for improvement and are desirable (in the light of prior knowledge and expert opinion) in addition to technically achievable.

Keywords DEA · Assurance region (AR) · Benchmarking · Closest targets

7.1 Introduction

DEA, as introduced in Charnes et al. (1978), is a methodology for the evaluation of relative efficiency of decision making units (DMUs) involved in production processes. The basic DEA models consist of a couple of dual problems, the so-called

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

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_7

multiplier and envelopment formulations. The multiplier model is the linear program equivalent to a fractional problem which provides an assessment of relative efficiency of the DMUs in terms of the classical efficiency ratios. The envelopment formulation carries out such assessment as the result of an evaluation of the DMUs within a technology which is constructed from the observations by assuming some postulates.

DEA, specifically through the envelopment models, has been widely used as a tool for the benchmarking with the purpose of improving performance of the DMUs. Cook et al. (2014) claim that *In the circumstance of benchmarking, the efficient DMUs, as defined by DEA, may not necessarily form a "production frontier", but rather lead to a "best-practice frontier*". The points on the best practice frontier allow us to identify benchmark performances for the inefficient units, while the targets are actually the coordinates of these benchmarks and represent levels of operation for the inefficient DMUs that would make them perform efficiently. As stated in Thanassoulis et al. (2008), in many practical applications one is more interested in determining targets that render the DMUs efficient than in determining their level of inefficiency. See Adler et al. (2013), Dai and Kuosmanen (2014) and Hung et al. (2010) for some recent references on applications of DEA and benchmarking.

The DEA models allow us to incorporate value judgments into the analysis through the addition of weight restrictions to the multiplier formulations. Some approaches to deal with this issue include the AR (Assurance Region) constraints (Thompson et al. 1986) and the cone-ratio models (Charnes et al. 1990). See Allen et al. (1997), Thanassoulis et al. (2004) and Cooper et al. (2011b) for some surveys on DEA models with weight restrictions. See also Lins et al. (2007), Portela et al. (2012) and Zanella et al. (2013) for applications of the radial CCR DEA model with weight restrictions. Since the weights attached by the model represent a relative value system of the inputs and outputs, weight restrictions make it possible to incorporate into the DEA models managerial preferences or prior views about the relative worth of inputs and outputs. For example, in the comparison of university departments in Beasley (1990) weight restrictions were used 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. Likewise, in the assessment of effectiveness of basketball players in Cooper et al. (2009) the weight restrictions used reflected value judgments from team coaches like "the importance attached to rebounds for playmakers should not be greater than that attached to field goal".

However, as Podinovski (2004, 2007a, b, 2015) has shown, DEA models with weight restrictions can also be seen under the perspective of models that allow us to incorporate technological judgments into the analysis. Specifically, as this author states, weight restrictions can be interpreted as production trade-offs. Under that perspective, restricted DEA models can be used for an evaluation of DMUs that incorporates information regarding certain simultaneous changes to the inputs and outputs which are technologically possible in the production process at all units. For instance, in the example carried out over some (hypothetical) university

departments in Podinovski (2007a) the following judgment was considered: "one researcher can be substituted by five teaching staff, without any detriment to the outputs (including publications)". This production trade-off was incorporated into the analysis by imposing the weight attached to research staff to be no greater than 5 times the weight attached to teaching staff in the multiplier model. In the application to agricultural farms in Turkish regions in Atici and Podinovski (2015), some judgments like "Any farm in the region can produce at least 0.75 t of barley instead of 1 t of wheat, without claiming additional resources" are elicited. This is translated into a weight restriction in which the weight attached to wheat is imposed to be greater than or equal to 0.75 times the weight attached to barley.

This chapter discusses the use of DEA models with weight restrictions for purposes of benchmarking and target setting. The developments are made by distinguishing between whether weight restrictions are used to incorporate technological judgments or value judgments. The implications of each of these two approaches in the models used for the benchmarking are analyzed. The basic restricted DEA models are based on a pre-emptive priority given to either the radial efficiency score (in the radial case) or the standard slacks (in the non-radial case), which may lead to targets outside the PPS. This is why we claim that these models are particularly useful when weight restrictions represent production trade-offs and restricted models are regarded as a mean to add new facets to the original frontier and/or extend the existing ones. In that case, targets derived from projections on to those facets can be considered as long as they reflect acceptable trade-offs between inputs and outputs. However, if weight restrictions are only a way of incorporating value judgments or preferences, then there may be no reason to argue that targets outside the PPS are attainable. It should be noted that the basic restricted DEA models may yield targets within the PPS, after some adjustments of the former projection which is determined by the objective function of those models are made. Nevertheless, it is shown that such approach can be enhanced, in particular through a more appropriate selection of benchmarks.

We develop here some new restricted DEA models that exploit the duality relations determined by the weight restrictions and provide targets (1) which are attainable (within the PPS), (2) which are Pareto-efficient and (3) which result from a selection of benchmarks that is made following a suitability criterion. Specifically, the models proposed find the closest targets. Minimizing the gap between actual and efficient performances seeks to find the most similar benchmarks to the unit that is being evaluated. This approach is therefore particularly useful when weight restrictions reflect value judgments and targets within the PPS must be ensured. In that case, it makes it possible to identify best practices and set targets which show the way for improvement with less effort and which are not only technically achievable but also consistent with the prior views of experts.

The chapter is organized as follows: Sect. 7.2 reviews the restricted formulations corresponding to the basic CCR DEA model. In Sect. 7.3 we briefly describe the approach for the benchmarking and target setting based on DEA models including weight restrictions that represent production trade-offs. In Sect. 7.4 we develop

some new restricted DEA models that ensure targets within the PPS. Section 7.5 illustrates the proposed approach with an empirical example. Last section concludes.

7.2 The Basic Restricted DEA Models

Throughout the paper, we suppose that we have n DMUs which use m inputs to produce s outputs. These are denoted by $(X_j, Y_j), j = 1, \ldots, n$. It is assumed that $X_j = (x_{1j}, \ldots, x_{mj})' \geq 0, \quad X_j \neq 0, j = 1, \ldots, n, \quad \text{and} \quad Y_j = (y_{1j}, \ldots, y_{sj})' \geq 0, \; Y_j \neq 0, \; j = 1, \ldots, n$. We also assume a DEA constant returns to scale (CRS) technology for the efficiency analysis and the benchmarking. Thus, the production possibility set (PPS), $T = \{(X,Y)/X \text{ can produce } Y\}$, can therefore be characterized as follows $T = \left\{(X,Y)/X \geq \sum_{j=1}^{n} \lambda_j X_j, Y \leq \sum_{j=1}^{n} \lambda_j Y_j, \lambda_j \geq 0\right\}$. The developments in the proposed approach can be straightforwardly extended to the variable returns to scale (VRS) case (Banker et al. 1984).

The CCR model (Charnes et al. 1978) generalized the single-output/input ratio measure of efficiency for a single DMU_0 in terms of a fractional linear programming formulation transforming the multiple output/input characterization of each DMU to that of a single virtual output and virtual input. That formulation is the following

$$\begin{array}{ll} \text{Max} & \dfrac{u'Y_0}{v'X_0} \\ \text{s.t.:} & \dfrac{u'Y_j}{v'X_j} \leq 1 \qquad j = 1, \ldots, n \\ & v \geq 0_m, u \geq 0_s \end{array} \tag{7.1}$$

The optimal value of (7.1) provides a measure of relative efficiency of DMU_0 as the ratio of a weighted sum of outputs to a weighted sum of inputs where the weights are selected trying to show DMU_0 in its best possible light subject to the constraint that no DMU can have a relative efficiency score greater than unity. By using the results in Charnes and Cooper (1962), model can be transformed into the following linear problem[1]

[1]We note that in (7.2) and (7.3) it suffices to consider the set E of extreme efficient DMUs. See Charnes et al. (1986) for the classification of DMUs into the sets E, E', F, NE, NE' and NF. This paper includes a procedure that allows to differentiating between the DMUs in E and E'.

$$\begin{array}{ll} \text{Max} & u'Y_0 \\ \text{s.t.}: & \\ & v'X_0 = 1 \\ & -v'X_j + u'Y_j \le 0 \quad j \in E \\ & v \ge 0_m, u \ge 0_s \end{array} \tag{7.2}$$

whose dual problem is

$$\begin{array}{ll} \text{Min} & \theta_0 \\ \text{s.t.}: & \\ & \sum_{j\in E} \lambda_j X_j \le \theta_0 X_0 \\ & \sum_{j\in E} \lambda_j Y_j \ge Y_0 \\ & \lambda_j \ge 0 \qquad j \in E \end{array} \tag{7.3}$$

Models (7.2) and (7.3) are a couple of dual problems which are known as the multiplier and envelopment formulations, respectively, of the CCR model. Actually, they correspond to the version input-oriented of that model, which is the one we use here for the developments (an output-oriented model could have been similarly used).

As said in the introduction, there exist different approaches for the addition of weight restrictions to the dual multiplier formulation of the used DEA models. Here, we deal with AR-I type restrictions (Thompson et al. 1986) like the ones below

$$\begin{aligned} L^I_{ii'} \le \frac{v_{i'}}{v_i} \le U^I_{ii'}, \quad i,i' = 1,\ldots,m, \quad i < i' \\ L^O_{rr'} \le \frac{u_{r'}}{u_r} \le U^O_{rr'}, \quad r,r' = 1,\ldots,s, \quad r < r' \end{aligned} \tag{7.4}$$

$L^I_{ii'}, U^I_{ii'}, L^O_{rr'}, U^O_{rr'}$ being some weight bounds.

The CCR DEA model is probably the one mostly used in practice when weight restrictions are considered. The dual multiplier formulation of the CCR-AR model can be formulated as follows

$$\begin{array}{ll} \text{Max} & u'Y_0 \\ \text{s.t.}: & \\ & v'X_0 = 1 \\ & -v'X_j + u'Y_j \le 0 \quad j \in E \\ & P'v \le 0_{2p} \\ & Q'u \le 0_{2q} \\ & v \ge 0_m, u \ge 0_s \end{array} \tag{7.5}$$

where

$$P' = \begin{pmatrix} L^{I}_{12} & -1 & 0 & \cdots & 0 & 0 \\ -U^{I}_{12} & 1 & 0 & \cdots & 0 & 0 \\ L^{I}_{13} & 0 & -1 & \cdots & 0 & 0 \\ -U^{I}_{13} & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ L^{I}_{1m} & 0 & 0 & \cdots & 0 & -1 \\ -U^{I}_{1m} & 0 & 0 & \cdots & 0 & 1 \\ 0 & L^{I}_{23} & -1 & \cdots & 0 & 0 \\ 0 & -U^{I}_{23} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & L^{I}_{2m} & 0 & \cdots & 0 & -1 \\ 0 & -U^{I}_{2m} & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & L^{I}_{m-1,m} & -1 \\ 0 & 0 & 0 & \cdots & -U^{I}_{m-1,m} & 1 \end{pmatrix}_{2pxm} \tag{7.6}$$

p being the number of pairs of inputs, that is, $p = \frac{m(m-1)}{2}$, and

$$Q' = \begin{pmatrix} L^{O}_{12} & -1 & 0 & \cdots & 0 & 0 \\ -U^{O}_{12} & 1 & 0 & \cdots & 0 & 0 \\ L^{O}_{13} & 0 & -1 & \cdots & 0 & 0 \\ -U^{O}_{13} & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ L^{O}_{1s} & 0 & 0 & \cdots & 0 & -1 \\ -U^{O}_{1s} & 0 & 0 & \cdots & 0 & 1 \\ 0 & L^{O}_{23} & -1 & \cdots & 0 & 0 \\ 0 & -U^{O}_{23} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & L^{O}_{2s} & 0 & \cdots & 0 & -1 \\ 0 & -U^{O}_{2s} & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & L^{O}_{s-1,s} & -1 \\ 0 & 0 & 0 & \cdots & -U^{O}_{s-1,s} & 1 \end{pmatrix}_{2qxs} \tag{7.7}$$

where $q = \frac{S(S-1)}{2}$. In case there is no restrictions involving v_i and $v_{i'}$ the corresponding rows in P' should be removed; we proceed similarly with the restrictions on the output weights.

The dual problem to the AR model (7.5) is the following envelopment formulation

$$\begin{array}{ll} \text{Min} & \theta_0 \\ \text{s.t. :} & \\ & \sum_{j\in E} \lambda_j X_j - P\pi \leq \theta_0 X_0 \\ & \sum_{j\in E} \lambda_j Y_j + Q\tau \geq Y_0 \\ & \lambda_j \geq 0, \pi \geq 0_{2p}, \tau \geq 0_{2q} \end{array} \tag{7.8}$$

where $\pi = \left(\pi_{12}^{-}, \pi_{12}^{+}, \ldots, \pi_{m-1,m}^{-}, \pi_{m-1,m}^{+}\right)'$ and $\tau = \left(\tau_{12}^{-}, \tau_{12}^{+}, \ldots, \tau_{s-1,s}^{-}, \tau_{s-1,s}^{+}\right)'$. Note that for each pair of restrictions on the input weights in (7.4) we have a couple of non-negative variables $\pi_{ii'}^{-}$ and $\pi_{ii'}^{+}$, i, i′ = 1, …, m, i < i′, by duality. The same happens with the restrictions on the output weights and their dual variables $\tau_{rr'}^{-}$ and $\tau_{rr'}^{+}$, r, r′ = 1, …, s, r < r′.

7.3 Weight Restrictions and Technological Judgments

Podinovski (2004, 2007a, b, 2015) has developed an approach in which weight restrictions in DEA models are interpreted as production trade-offs. Under that perspective, we can incorporate into the analysis technological judgments regarding simultaneous changes to the inputs and outputs that are technologically possible in the production process. The incorporation of weight restrictions which represent production trade-offs leads to models that carry out the benchmarking of the DMUs within an extended PPS, whose frontier results from the addition of new facets to the original efficient frontier and/or the extension of some of the existing ones. This is illustrated in the following numerical example.

7.3.1 Numerical Example

Consider the situation in which we have the following DMUs which use 2 inputs to produce 1 constant output (Table 7.1).

Table 7.1 Data of numerical example

	DMU A	DMU B	DMU C	DMU D	DMU E
x_1	4	5	10	14	9
x_2	6	4	2	1.5	5
y	10	10	10	10	10

For the evaluation of these DMUs we are willing to make the following assumptions on admissible trade-offs between the two inputs:

Assumption 1 All the DMUs can maintain their production of outputs if x_1 is reduced by 5 units provided that x_2 is increased by at least 3 units (in certain proportion).

Assumption 2 All the DMUs can maintain their production of outputs if x_2 is reduced by 1 unit provided that x_1 is increased by at least 4 units (in certain proportion).

The incorporation of such information regarding admissible trade-offs between the inputs leads to the following expansion of the original PPS

$$\lambda_A \begin{pmatrix} 4 \\ 6 \end{pmatrix} + \lambda_B \begin{pmatrix} 5 \\ 4 \end{pmatrix} + \lambda_C \begin{pmatrix} 10 \\ 2 \end{pmatrix} + \lambda_D \begin{pmatrix} 14 \\ 1.5 \end{pmatrix} - \pi_{12}^{-} \begin{pmatrix} 5/3 \\ -1 \end{pmatrix} - \pi_{12}^{+} \begin{pmatrix} -4 \\ 1 \end{pmatrix} \leq \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$$10\lambda_A + 10\lambda_B + 10\lambda_C + 10\lambda_D \geq Y$$

$$\lambda_A, \lambda_B, \lambda_C, \lambda_D, \pi_{12}^{-}, \pi_{12}^{+} \geq 0$$

provided that $X_1,\ X_2,\ Y \geq 0$ (see Podinovski 2015).

In Fig. 7.1 we can see, at level Y = 10, the efficient frontier of the original PPS (the solid line connecting points A, B, C and D) and the two new facets that are added (the dashed lines) as the result of considering the information on trade-offs. These two facets and the segment $\overline{BC}$, which is actually the AR frontier of the original PPS, form the frontier of the expanded PPS. Note that DMUs A and D, which are technically efficient, become inefficient.

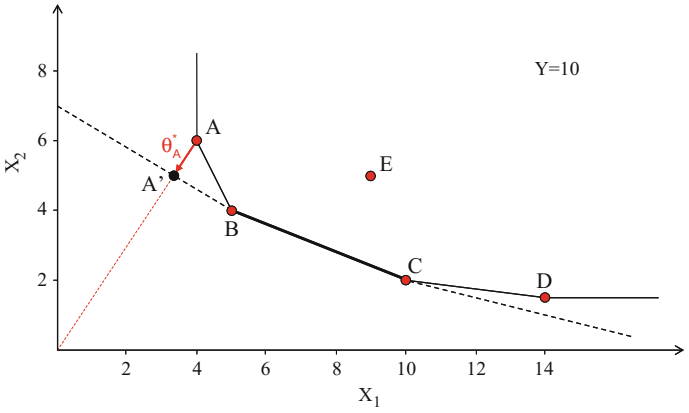


Fig. 7.1 Extended PPS and radial targets (DMU A)

The evaluation of efficiency of any of the DMUs, for example DMU A, can be made by solving the following envelopment model

$$\begin{array}{ll}\text{Min} & \theta_A \\ \text{s.t.:} & \lambda_A\begin{pmatrix}4\\6\end{pmatrix}+\lambda_B\begin{pmatrix}5\\4\end{pmatrix}+\lambda_C\begin{pmatrix}10\\2\end{pmatrix}+\lambda_D\begin{pmatrix}14\\1.5\end{pmatrix}-\pi_{12}^{-}\begin{pmatrix}5/3\\-1\end{pmatrix}-\pi_{12}^{+}\begin{pmatrix}-4\\1\end{pmatrix}\le\begin{pmatrix}\theta_A X_1\\ \theta_A X_2\end{pmatrix} \\ & 10\lambda_A+10\lambda_B+10\lambda_C+10\lambda_D\ge Y \\ & \lambda_A,\lambda_B,\lambda_C,\lambda_D,\pi_{12}^{-},\pi_{12}^{+}\ge 0\end{array} \tag{7.9}$$

Solving (7.9) leads to the projection point with inputs $\theta_A^*\begin{pmatrix}4\\6\end{pmatrix}$, where $\theta_A^*=\frac{5}{6}$, and output Y = 10. At level Y = 10, that projection corresponds to the point A′ in Fig. 7.1, and represents a benchmark for DMU A on one of the new facets added to the efficient frontier of the original PPS. Its coordinates provide attainable targets (radial targets) for that DMU because the extended PPS can be considered as a valid one as the result of the assumption on the trade-offs.

Note that, by duality, the optimal value of the model above can be obtained by solving the following multiplier model

$$\begin{array}{ll}\text{Max} & 10u \\ \text{s.t.:} & 4v_1+6v_2=1 \\ & -4v_1-6v_2+10u\le 0 \\ & -5v_1-4v_2+10u\le 0 \\ & -10v_1-2v_2+10u\le 0 \\ & -14v_1-1.5v_2+10u\le 0 \\ & \frac{5}{3}\le\frac{v_2}{v_1}\le 4 \\ & v_1,v_2,u\ge 0\end{array} \tag{7.10}$$

which includes the information associated with trade-offs by means of the weight restrictions $(v_1 \quad v_2)\begin{pmatrix}5/3\\-1\end{pmatrix}\le 0$ and $(v_1 \quad v_2)\begin{pmatrix}-4\\1\end{pmatrix}\le 0$. Thus, we can see that, when we want to incorporate information regarding production trade-offs, the efficiency of the DMUs can be equivalently evaluated by using weight restrictions in the multiplier formulation of the DEA models.

In general, when weight restrictions are interpreted as production trade-offs, models (7.5) and (7.8) provide measures of efficiency of the DMUs that take into consideration the corresponding technological judgments. Model (7.8), in particular, provides targets (radial targets), perhaps outside the original PPS, in terms of its optimal solutions as

$$\begin{aligned}\hat{X}_0 &= \theta_0^* X_0 \quad \left(\geq \sum_{j\in E} \lambda_j^* X_j - P\pi^*\right)\\ \hat{Y}_0 &= Y_0 \quad \left(\leq \sum_{j\in E} \lambda_j^* Y_j + Q\tau^*\right)\end{aligned} \tag{7.11}$$

It is worth mentioning that the targets (7.11) will not be efficient in the sense of Pareto if slacks are present when model (7.8) is solved. The literature offers several ways to deal with this issue. These are described in the next subsection.

7.3.2 *Pareto-Efficient Targets*

Performing a second stage which maximizes the slacks starting from the radial projection is one of the ways commonly used to find Pareto-efficient targets. However, as Podinovski (2007b) shows, a standard second stage applied to models with weight restrictions that represent production trade-offs may result in benchmarks with meaningless negative values of some inputs. This is why this author has proposed a corrected procedure of the conventional second stage which ensures non-negativity of the variables (see that paper for details). In any case, note that the targets that are found in a second stage would not necessarily preserve the mix of inputs.

As an alternative approach for finding Pareto-efficient targets, the use from the beginning of a non-radial model could be considered. Thanassoulis et al. (2008) state that non-radial DEA models are the appropriate instrument for the setting of targets and the benchmarking. In particular, these models ensure that the identified targets lie on the Pareto-efficient subset of the frontier. The following additive-type model is an envelopment formulation that includes the terms $P\pi$ and $Q\tau$ (sometimes called residues in the literature) which, by duality, are associated with the weight restrictions (7.4)

$$\begin{aligned}\text{Max} \quad & Z_0 = 1'_m s^- + 1'_s s^+ \\ \text{s.t. :} \quad & \\ & \sum_{j\in E} \lambda_j X_j - P\pi = X_0 - s^- \qquad (12.1)\\ & \sum_{j\in E} \lambda_j Y_j + Q\tau = Y_0 + s^+ \qquad (12.2)\\ & \lambda_j \geq 0, s^- \geq 0_m, s^+ \geq 0_s, \pi \geq 0_{2p}, \tau \geq 0_{2q}\end{aligned} \tag{7.12}$$

A VRS formulation of (7.12) can be found in Charnes et al. (1994) (Chap. 3, p. 56), which is actually the restricted version of the additive model (Charnes et al. 1985). For ease of exposition, non-radial models in this chapter are formulated in terms of the L_1-norm, albeit the developments can be straightforwardly adapted to deal with the weighted additive models (Lovell and Pastor 1995). These include the invariant additive model (Charnes et al. 1987), the RAM model (Cooper et al. 1999)

and the BAM models (Cooper et al. 2011a). In fact, as Thrall (2000) points out, the weights attached to the slacks in the objective function of the additive models can represent an additional way of incorporating value judgments.

Like (7.8) and (7.12) may yield targets (perhaps outside the original PPS) in terms of its optimal solutions as follows

$$\begin{aligned}\hat{X}_0 &= X_0 - s^{-*} \ (= \sum_{j\in E} \lambda_j^* X_j - P\pi^*) \\ \hat{Y}_0 &= Y_0 + s^{+*} \ (= \sum_{j\in E} \lambda_j^* Y_j + Q\tau^*)\end{aligned} \tag{7.13}$$

which are obviously Pareto-efficient.

It should be noted that, as it happens with the conventional second stage used with the radial models, we may have some negative values for the targets of the inputs in (7.13). In any event, one of the main drawbacks of these approaches is that the slacks, both in the second stage procedures and in (7.12), are maximized (which guarantees that the efficient frontier is reached), while they should be minimized in order to find the most similar benchmark to the DMU_0 under evaluation. As already said, this would show DMU_0 the way for improvement with less effort.

7.4 Weight Restrictions and Value Judgments

In this section, we develop some new restricted DEA models that allow us to find targets which lie within the original PPS. These models will be particularly useful for the benchmarking of DMUs when weight restrictions are utilized to incorporate preferences or value judgments into the analysis. In those cases, the targets provided by the basic restricted DEA models, both radial (7.11) and non-radial (7.13), may be deemed unacceptable, because it cannot be ensured that they are determined by projection points which belong to the original PPS, in which case there would be no reason to argue that they are attainable.

It should be noted that the standard restricted DEA models may also yield targets lying within the PPS by using their optimal solutions. In the radial case, these targets can be obtained by using the optimal solutions of (7.8) as follows

$$\begin{aligned}\hat{X}_0 &= \theta_0^* X_0 + P\pi^* \quad (\geq \sum_{j\in E} \lambda_j^* X_j) \\ \hat{Y}_0 &= Y_0 - Q\tau^* \quad (\leq \sum_{j\in E} \lambda_j^* Y_j)\end{aligned} \tag{7.14}$$

or, alternatively, by using the following formulae that are proposed in Cooper et al. (2007) $\hat{X}_0 = \theta_0^* X_0 - s^{-*} + P\pi^* \ (= \sum_{j\in E} \lambda_j^* X_j)$ and $\hat{Y}_0 = Y_0 + s^{+*} - Q\tau^* \ (= \sum_{j\in E} \lambda_j^* Y_j)$, which take into account the slacks at optimum, thus providing targets Pareto-efficient.

Turning to the numerical example above, Fig. 7.2 depicts the benchmarking of DMU A according to (7.14). It can be seen that, after the adjustments determined by the residues $P\pi^*$ and $Q\tau^*$ are made, we eventually have a referent (DMU B) for that unit on the frontier of the original PPS. Specifically, the radial projection $A' = \frac{5}{6}\begin{pmatrix}4\\6\end{pmatrix}$ is adjusted taking into account that $\pi_{12}^{-*} = 1$, so that $\frac{5}{6}\begin{pmatrix}4\\6\end{pmatrix} + 1 \times \begin{pmatrix}5/3\\-1\end{pmatrix} = \begin{pmatrix}5\\4\end{pmatrix}$, which is the input vector of DMU B, can be set as targets for DMU A on the AR efficient frontier of the original PPS.

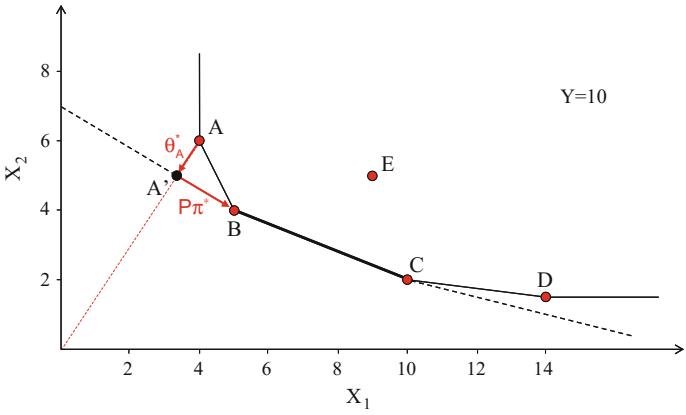


Fig. 7.2 Targets within the PPS (radial models)

Note, however, that this target setting involves changes in the input mix of DMU A. In fact, Thanassoulis et al. (2008) claim that targets from restricted models lying on the original PPS frontier will generally be non-radial targets. Therefore, it makes little sense to use model (7.8) for the benchmarking, which is a model that gives pre-emptive priority to the radial contraction of the inputs of DMU_0, if one is eventually interested in targets within the PPS, because that model does not ensure the preservation of the mix. In addition, it could be argued that the targets (7.14) are to some extent arbitrarily derived. Since the residues $P\pi$ and $Q\tau$, which represent the adjustments that should be made to reach the PPS, are not considered in the objective, then the ultimate benchmark would be actually found without following any criterion of optimality that relates the unit being evaluated and its targets.

Similar difficulties are found when the non-radial model (7.12) is used. Targets within the PPS can be obtained by using its optimal solutions as follows

$$\begin{aligned}\hat{X}_0 &= X_0 - s^{-*} + P\pi^* \left(= \sum_{j \in E} \lambda_j^* X_j\right) \\ \hat{Y}_0 &= Y_0 + s^{+*} - Q\tau^* \left(= \sum_{j \in E} \lambda_j^* Y_j\right)\end{aligned} \tag{7.15}$$

However, model (7.12) is based on a pre-emptive priority given to the standard slacks and ignores the residues in the objective (as in the radial case). Thus, the adjustments needed to reach the original PPS are not considered in the optimization that is made, and so, it can be argued again that the resulting targets are to some extent arbitrarily set. In addition, the slacks in (7.12) are maximized, when they should be minimized in order to find the closest targets.

Bearing in mind the above, we develop an approach for the benchmarking and the setting of targets with restricted DEA models taking into account the following considerations:

- It is an approach to be used when targets within the PPS must be ensured. That could be the case if, for example, weight restrictions were regarded as a way of incorporating preferences or value judgments and there are no reasons to consider targets outside the PPS.
- It is an approach based on non-radial models. As said before, additive-type models ensure Pareto-efficient targets. In addition, as Thanassoulis et al. (2008) claim, non-radial targets (within the original PPS frontier) may be preferable under restricted models if one is not certain about the trade-offs implicit in the new facets added to the PPS (as the result of the addition of weight restrictions).
- The identification of benchmarks and the setting of targets with this approach are based on a criterion of optimality. Specifically, we look for the closest targets. In general, minimizing the gap between actual and efficient performances allows us to identify the most similar efficient benchmark to the unit under assessment and, consequently, ensures the achievement of improvement with less effort. Thus, if weight restrictions are used to reflect value judgments, the model we propose identifies best practices which are in line with the prior knowledge and expert opinion and show the unit being evaluated the easiest way for improvement.

The weight restrictions (7.4) give rise by duality to the following constraints in the primal envelopment formulation of the additive-type DEA models: $\sum_{j \in E} \lambda_j X_j = X_0 + P\pi - s^-$ and $\sum_{j \in E} \lambda_j Y_j = Y_0 - Q\tau + s^+$, for some $\lambda_j \geq 0, s^- \geq 0_m, s^+ \geq 0_s, \pi \geq 0_{2p}, \tau \geq 0_{2q}$. These constraints can be seen as determining, through the standard slacks and the residues, the movements within the PPS that DMU_0 may follow in its search for a benchmark. The points $(X_0 + P\pi, Y_0 - Q\tau)$ result from substitutions between the inputs of DMU_0 $(P\pi)$ and/or substitutions between its outputs $(Q\tau)$, and $\left(X_0 + P\pi - s^-, Y_0 - Q\tau + s^+\right)$ are therefore points within the PPS that dominate them. Note that in the evaluation of a given DMU_0 with unrestricted models only the points of the PPS that dominate this unit are considered as potential benchmarks. In absence of any other

information, these are the only ones that can be deemed to perform better. Non-dominating points would perform better than DMU_0 for some weights but not for others. However, if some information on the worth of inputs and outputs is available, other production plans (in addition to the points that dominate DMU_0) can be considered for the benchmarking. Specifically, the points $\left(X_0 + P\pi - s^-, Y_0 - Q\tau + s^+\right)$ within the PPS may include other plans (aside from the dominating points) that can be used as benchmarks for that unit because they represent better performances according to allowable weights. This is shown graphically with the data of the numerical example.

7.4.1 Numerical Example (Cont.)

Consider again the data of the numerical example, where the DMUs are evaluated taking into account the weight restrictions $5/3 \leq v_2/v_1 \leq 4$. Now, we suppose that these weight restrictions are used as way to incorporate information regarding the relative importance to be attached to the inputs in the evaluation of the DMUs. Therefore, there is no reason to consider targets outside the PPS.

As an effect of considering the residues associated by duality with the weight restrictions, the selection of potential benchmarks for a given DMU will include points of the PPS which result from both (1) increasing x_1 by 5 units, provided that x_2 is reduced by at least 3 units (in certain proportion) and (2) increasing x_2 by 1 unit, provided that x_1 is reduced by at least 4 units (in certain proportion), aside from dominating points.

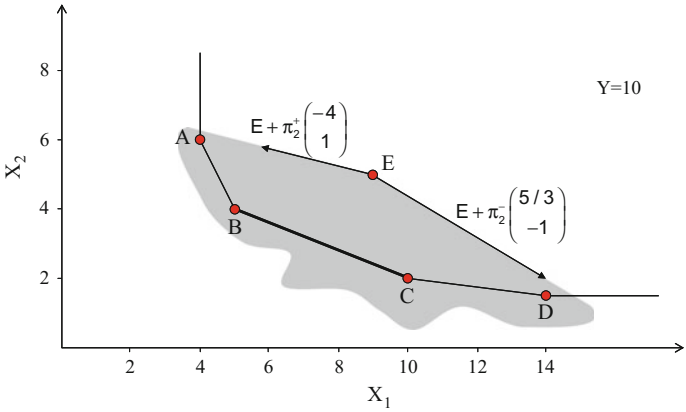


Fig. 7.3 Selection of benchmarks within the PPS (DMU E)

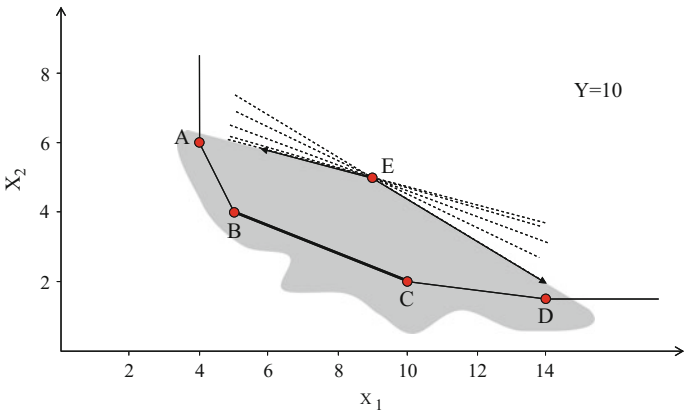


Fig. 7.4 Virtual input isolines for weights satisfying the weight restrictions (DMU E)

Figure 7.3 depicts the selection of benchmarks for DMU E with this approach. The shadow area of that figure corresponds to the points $E+\pi_{12}^{-}\begin{pmatrix}5/3\\-1\end{pmatrix}+\pi_{12}^{+}\begin{pmatrix}-4\\1\end{pmatrix}-\begin{pmatrix}s_1^{-}\\s_2^{-}\end{pmatrix}$, $\pi_{12}^{-},\pi_{12}^{+},s_1^{-},s_2^{-},\lambda_j\geq 0$ (at level Y = 10). Thus, the points of the PPS in that area are, in principle, potential benchmarks for DMU E, because they represent production plans that outperform the one of that unit. Note that the value of the virtual input $v_1x_1+v_2x_2$ in those points, for weights (v_1,v_2,u) satisfying the weight restrictions, is always no lower than that in E. This is shown in Fig. 7.4, where we have represented the family of lines $v_1x_1+v_2x_2=c$ which result from varying those input weights, c being the corresponding virtual input in DMU E. Eventually, the selection of benchmarks should be made ensuring that the referents chosen among those potential benchmarks are, at least, AR-efficient.

If we proceed in a similar manner as with DMU E, we can see that the residues and the standard slacks determine no other potential benchmarks for DMUs B and C aside from themselves (see in Fig. 7.5 that the shadow areas do not include points of the PPS; except, obviously, the input vectors of those two units). This is why they are AR-efficient.

The case of DMU A is particularly interesting because it is a unit technically efficient which becomes inefficient as the result of the weight restrictions. The selection of benchmarks for that unit is depicted in Fig. 7.6. We can see that there are other production plans within the PPS that outperform DMU A, in particular those in the segment determined by DMUs B and C, which are AR-efficient. Therefore, they could be considered as its referents, albeit an additional selection criterion should be used in order to make the choice of the ultimate benchmark among them.

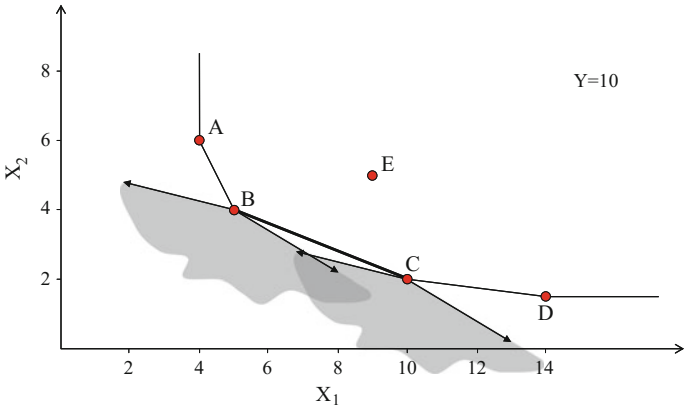


Fig. 7.5 Selection of benchmarks within the PPS (DMUs B and C)

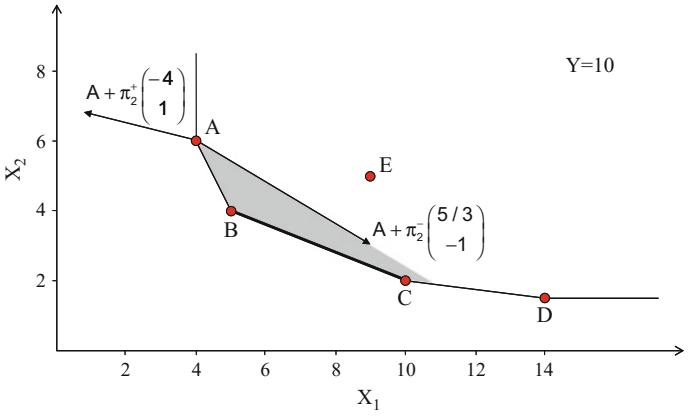


Fig. 7.6 Selection of benchmarks within the PPS (DMU E)

A similar situation is found in the evaluation of DMU D (see Fig. 7.7). However, it should be pointed out that in that case some of the points of the AR efficient frontier are not so far considered as potential benchmarks for DMU D. Specifically, the constraints in the envelopment models associated with the weight restrictions determine that, among the points on the AR efficient frontier, only the points in segment $\overline{D'C}$ are considered for the benchmarking of DMU D. In particular, DMU B cannot be a referent for DMU D, because we cannot state that DMU B outperforms DMU D, as this depends on the weights that are used. For example, if the weights chosen are (6, 10, 7), then it can be said that DMU B performs better than DMU D (the virtual input of DMU B would be 70 versus 99 in the case of DMU D); however, if the weights are (5, 20, 9) then it is the other way around (the

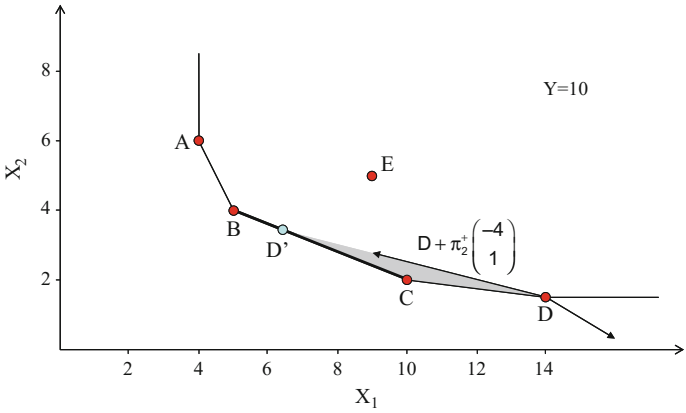


Fig. 7.7 Selection of benchmarks within the PPS (DMU D)

virtual input of DMU B would be 105 versus 100 in the case of DMU D). Note that both weight vectors satisfy the weight restrictions.

7.4.2 A Model to Find the Closest Targets

The analysis of the numerical example made above has shown that the envelopment models which result by duality from the multiplier formulations including weight restrictions allow us to identify potential benchmarks for the DMU_0 in the sense that they represent better performances. However, the selection of the ultimate benchmark should be made ensuring that it corresponds to a production plan which is consistent with the prior knowledge and the views of experts, which is efficient in the sense of Pareto and, if possible, which is chosen following some suitability criterion.

Concerning the suitability criterion, the approach we propose here is intended to find a benchmark which represents the most similar performance to that of DMU_0. To achieve that, we should minimize the total deviations between actual inputs/outputs, (X_0, Y_0), and targets, $\left(X_0 + P\pi - s^-, Y_0 - Q\tau + s^+\right)$, making sure that the latter point (the projection point) belongs to the AR Pareto-efficient frontier of the PPS.

The model which provides the targets that are wanted is the following problem

$$
\begin{aligned}
&\text{Min} \quad z_0 = \left\|s^- - P\pi\right\|_1 + \left\|s^+ - Q\tau\right\|_1 \\
&\text{s.t. :} \\
&\qquad \sum_{j\in E} \lambda_j X_j = X_0 - (s^- - P\pi) && (16.1)\\
&\qquad \sum_{j\in E} \lambda_j Y_j = Y_0 + (s^+ - Q\tau) && (16.2)\\
&\qquad -v'X_j + u'Y_j + d_j = 0 \qquad j \in E && (16.3)\\
&\qquad P'v \le 0_{2p} && (16.4)\\
&\qquad Q'u \le 0_{2q} && (16.5)\\
&\qquad v \ge 1_m && (16.6)\\
&\qquad u \ge 1_s && (16.7)\\
&\qquad d_j \le M\, b_j \qquad j \in E && (16.8)\\
&\qquad \lambda_j \le M(1 - b_j) \qquad j \in E && (16.9)\\
&\qquad \lambda_j, d_j \ge 0, b_j \in \{0,1\} \qquad j \in E \\
&\qquad s^- \ge 0_m, s^+ \ge 0_s, \pi \ge 0_{2p}, \tau \ge 0_{2q}
\end{aligned}
\tag{7.16}
$$

where

1. $\left\|s^- - P\pi\right\|_1 = \left|s_1^- - \sum_{i'=2}^{m}\left(L^I_{1i'}\pi^-_{1i'} - U^I_{1i'}\pi^+_{1i'}\right)\right| + \sum_{i=2}^{m-1}\left|s_i^- - \sum_{k=1}^{i-1}\left(-\pi^-_{ki} + \pi^+_{ki}\right) - \sum_{i'=i+1}^{m}\left(L^I_{ii'}\pi^-_{ii'} - U^I_{ii'}\pi^+_{ii'}\right)\right| + \left|s_m^- - \sum_{k=1}^{m-1}\left(-\pi^-_{km} + \pi^+_{km}\right)\right|$. That is, the L_1-norm of the deviation vector between actual inputs and input targets.
2. $\left\|s^+ - Q\tau\right\|_1 = \left|s_1^+ - \sum_{r'=2}^{s}\left(L^O_{1r'}\tau^-_{1r'} - U^O_{1r'}\tau^+_{1r'}\right)\right| + \sum_{r=2}^{s-1}\left|s_r^+ - \sum_{k=1}^{r-1}\left(-\tau^-_{kr} + \tau^+_{kr}\right) - \sum_{r'=r+1}^{s}\left(L^O_{rr'}\tau^-_{rr'} - U^O_{rr'}\tau^+_{rr'}\right)\right| + \left|s_s^+ - \sum_{k=1}^{s-1}\left(-\tau^-_{ks} + \tau^+_{ks}\right)\right|$, the L_1-norm of the deviation vector between actual outputs and output targets. And
3. M is a big positive quantity.

Model (7.16) is in line with others that have already been proposed to minimize the distance to the efficient frontier of the PPS (see, for example, Aparicio et al. (2007) and Ruiz et al. (2015)[2]). It is important to highlight that, in minimizing instead of maximizing the deviations, we have to make sure that the projection points obtained lie on the Pareto-efficient frontier of the PPS, specifically the AR frontier. Note that maximizing the standard slacks in (7.12) ensures that the AR Pareto-efficient frontier of the PPS is reached. However, to make sure of that, when deviations are minimized, model (7.16) follows a primal-dual approach in the sense that it includes the constraints of both the envelopment and the multiplier formulations of the AR model considered [(7.16)–(16.2) and (16.3)–(16.7), respectively]. The key is in the restrictions (16.8) and (16.9), which link those two groups of constraints [(16.8) and (16.9) actually result from a characterization of the DEA Pareto efficient frontier; see Ruiz et al. 2015]. According to these restrictions, if

[2]Other papers dealing with closest targets in DEA include Portela et al. (2003), Tone (2010), Fukuyama et al. (2014), Aparicio and Pastor (2014) and Ruiz and Sirvent (2016).

$\lambda_j > 0$, then (16.9) implies that $d_j = 0$ by virtue of (16.8). Thus, if DMU_j participates actively as a referent in the evaluation of DMU_0, then it necessarily belongs to $-v'X + u'Y = 0$. That is, the benchmarks that model (7.16) finds for DMU_0 are combinations of DMU_j's in E that are all on a same facet of the Pareto AR efficient frontier, because those DMU_j's belong all to a supporting hyperplane of T whose coefficients are non-zero and satisfy the AR restrictions.

Eventually, targets on the AR Pareto-efficient frontier of the PPS can be derived by using the optimal solutions of (7.16) as follows

$$\begin{aligned} \hat{X}_0 &= X_0 - s^{-*} + P\pi^* \left(= \sum_{j \in E} \lambda_j^* X_j\right) \\ \hat{Y}_0 &= Y_0 + s^{+*} - Q\tau^* \left(= \sum_{j \in E} \lambda_j^* Y_j\right) \end{aligned} \tag{7.17}$$

That is, these targets formulae are the same as those in (7.15), but use the optimal solutions of a different model which imposes a suitable criterion for the selection of benchmarks.

7.4.3 Numerical Example (Cont.)

The use of model (7.16) with the DMUs of the numerical example leads to the following results. In the evaluation of DMU A, the feasible solutions are associated with all the points on the AR efficient frontier with Y ≥ 10 (at level Y = 10 this corresponds to the segment $\overline{BC}$). At optimum, $\pi_{12}^{-*} = 0.6$, $\pi_{12}^{+*} = 0$, $s_1^{-*} = 0$, $s_2^{-*} = 1.4$ and $s_1^{+*} = 0$. Thus, DMU B is selected as its referent, which is the closest point on the AR frontier. In the case of DMU D, model (7.16) only considers as potential benchmarks the points on the AR efficient frontier in segment $\overline{D'C}$ (see again Fig. 7.7). At optimum, $\pi_{12}^{-*} = 0$, $\pi_{12}^{+*} = 0.5$, $s_1^{-*} = 2$, $s_2^{-*} = 0$ and $s_1^{+*} = 0$, so DMU C is selected as its benchmark.

Remark 1 [The linearization of the objective of model (7.16)]
Model (7.16) is a non-linear problem as the result of using in its objective function the absolute values of the deviations between actual inputs/outputs and targets. However, this problem can be reformulated without the absolute values, as it is explained next. We introduce the new decision variables $\gamma_i^+, \gamma_i^- \geq 0$, $i = 1, \ldots, m$, $\delta_r^+, \delta_r^- \geq 0$, $r = 1, \ldots, s$, and add the restrictions $s_1^- - \sum_{i'=2}^{m} \left(L_{1i'}^I \pi_{1i'}^- - U_{1i'}^I \pi_{1i'}^+\right) = \gamma_1^+ - \gamma_1^-$, $s_i^- - \sum_{k=1}^{i-1} \left(-\pi_{ki}^- + \pi_{ki}^+\right) - \sum_{i'=i+1}^{m} \left(L_{ii'}^I \pi_{ii'}^- - U_{ii'}^I \pi_{ii'}^+\right) = \gamma_i^+ - \gamma_i^-$, $i = 2, \ldots, m-1$, $s_m^- - \sum_{k=1}^{m-1} \left(-\pi_{km}^- + \pi_{km}^+\right) = \gamma_m^+ - \gamma_m^-$ and $s_1^+ - \sum_{r'=2}^{s} \left(L_{1r'}^O \tau_{1r'}^- - U_{1r'}^O \tau_{1r'}^+\right) = \delta_1^+ - \delta_1^-$, $s_r^+ - \sum_{k=1}^{r-1} \left(-\tau_{kr}^- + \tau_{kr}^+\right) - \sum_{r'=r+1}^{s} \left(L_{rr'}^O \tau_{rr'}^- - U_{rr'}^O \tau_{rr'}^+\right) = \delta_r^+ - \delta_r^-$, $r = 2, \ldots, s-1$, $s_s^+ - \sum_{k=1}^{s-1} \left(-\tau_{ks}^- + \tau_{ks}^+\right) = \delta_s^+ - \delta_s^-$. Then, minimizing the non-linear objective in (7.16) is equivalent to minimizing the linear objective

function $\sum_{i=1}^{m}\left(\gamma_i^+ + \gamma_i^-\right) + \sum_{r=1}^{s}\left(\delta_r^+ + \delta_r^-\right)$ subject to the resulting set of constraints. Thus, (7.16) becomes a mixed-integer linear programming model.

Remark 2 Solving (7.16) in practice

The formulation of model (7.16) seeks that the DMU_j's in E that participate actively as a referent in the evaluation of DMU_0 necessarily belong to the same facet of the efficient frontier. This is actually achieved by means of the constraints (16.8) and (16.9), which include the classical big M and binary variables. Solving (7.16) in practice may involve setting a value for M, say 10^6, and this might become an issue because some projection points (closer to DMU_0) could remain unconsidered (unless a lower bound for M were found). Nevertheless, (7.16) can be solved by reformulating these constraints using Special Ordered Sets (SOS) (Beale and Tomlin 1970). SOS Type 1 is a set of variables where at most one variable may be nonzero. Therefore, if we remove (16.8) and (16.9) from the formulation and define instead a SOS Type 1, S_j, for each pair of variables $\{\lambda_j, d_j\}$, $j \in E$, then it is ensured that λ_j and d_j cannot be simultaneously positive for DMU_j's, $j \in E$. CPLEX Optimizer (and also LINGO) can solve LP problems with SOS by using branching strategies that take advantage of this type of variables.

Remark 3 Ruiz et al. (2015) also propose a non-radial model with weight restrictions which seeks to find the closest targets. In that paper, the constraints (16.1) and (16.2) are replaced by $\sum_{j\in E}\lambda_j X_j = X_0 - s^-$ and $\sum_{j\in E}\lambda_j Y_j = Y_0 + s^+$, where s^- and s^+ contain variables unrestricted in sign. That is, the total input and output deviations are considered as simply free variables. Thus, Ruiz et al. (2015) does not follow exactly a primal-dual approach because these constraints of the envelopment formulation are not the ones that would correspond by duality to those in the multiplier formulation which include the weight restrictions. As a result, the model proposed by those authors makes the search for the targets of DMU_0 in the whole AR Pareto-efficient frontier, without considering the requirement on the allowable substitutions (reallocations), $P\pi$ and $Q\tau$, which result from the weight restrictions. This can be observed in Fig. 7.7. As said before, at level Y = 10, model (7.16) only considers as potential benchmarks for DMU D the points on the AR efficient frontier in segment $\overline{D'C}$. However, the approach in Ruiz et al. (2015) would consider the segment $\overline{BC}$ (the whole AR frontier), for the benchmarking of that unit.

7.4.4 A Unit-Invariant Model

In practice, we might have to consider the units of measurement because both the objective function of (7.16) is an aggregation of deviations in inputs and outputs and the weight restrictions in that model are formulated in terms of absolute weights. The following problem is an invariant version of model (7.16)

$$
\begin{aligned}
\text{Min} \quad & z_0 = \left\|\tilde{s}^- - \tilde{P}\tilde{\pi}\right\|_1^{\omega_I} + \left\|\tilde{s}^+ - \tilde{Q}\tilde{\tau}\right\|_1^{\omega_O} \\
\text{s.t. :} \quad & \\
& \sum_{j\in E} \tilde{\lambda}_j X_j = X_0 - (\tilde{s}^- - \tilde{P}\tilde{\pi}) && (18.1)\\
& \sum_{j\in E} \tilde{\lambda}_j Y_j = Y_0 + (\tilde{s}^+ - \tilde{Q}\tilde{\tau}) && (18.2)\\
& -\tilde{v}'X_j + \tilde{u}'Y_j + \tilde{d}_j = 0 \qquad j \in E && (18.3)\\
& \tilde{P}'\tilde{v} \le 0_{2p} && (18.4)\\
& \tilde{Q}'\tilde{u} \le 0_{2q} && (18.5)\\
& \bar{X}\tilde{v} \ge 1_m && (18.6)\\
& \bar{Y}\tilde{u} \ge 1_s && (18.7)\\
& \tilde{d}_j \le M\,\tilde{b}_j \qquad j \in E && (18.8)\\
& \tilde{\lambda}_j \le M\,(1 - \tilde{b}_j) \qquad j \in E && (18.9)\\
& \tilde{\lambda}_j, \tilde{d}_j \ge 0, \tilde{b}_j \in \{0,1\} \qquad j \in E && \\
& \tilde{s}^- \ge 0_m, \tilde{s}^+ \ge 0_s, \tilde{\pi} \ge 0_{2p}, \tilde{\tau} \ge 0_{2q} &&
\end{aligned}
\tag{7.18}
$$

where

1. $\left\|\tilde{s}^- - \tilde{P}\tilde{\pi}\right\|_1^{\omega_I} = \left\|\bar{X}^{-1}\left(\tilde{s}^- - \tilde{P}\tilde{\pi}\right)\right\|_1$, $\bar{X}$ being the diagonal matrix having as entries the averages of the inputs, i.e., $\bar{X} = \text{diag}(\bar{x}_1, \ldots, \bar{x}_m)$. That is, we use now a weighted L_1-norm of the deviations between actual inputs and targets, the weights being the inverse of the input averages. Such specification of the weighted L_1-norm has already been used to weight the standard slacks in additive-type models (see Thrall 1996).
   Likewise, $\left\|\tilde{s}^+ - \tilde{Q}\tilde{\tau}\right\|_1^{\omega_O} = \left\|\bar{Y}^{-1}\left(\tilde{s}^+ - \tilde{Q}\tilde{\tau}\right)\right\|_1$, where $\bar{Y} = \text{diag}(\bar{y}_1, \ldots, \bar{y}_s)$.
2. $\tilde{P} = \bar{X}P$ and $\tilde{Q} = \bar{Y}Q$, which means that the weight restrictions (7.4) are now formulated as

$$
\begin{aligned}
& L^I_{ii'} \le \frac{v_{i'}\bar{x}_{i'}}{v_i\bar{x}_i} \le U^I_{ii'},\ i,i' = 1,\ldots,m,\ i < i' \\
& L^O_{rr'} \le \frac{u_{r'}\bar{y}_{r'}}{u_r\bar{y}_r} \le U^O_{rr'},\ r,r' = 1,\ldots,s,\ r < r'
\end{aligned}
\tag{7.19}
$$

That is, we consider weight restrictions on virtual input and outputs, specifically on an average DMU (see Wong and Beasley 1990).

It should be noted that if we make the following change of variables in (7.18): $s^- = \bar{X}^{-1}\tilde{s}^-$, $s^+ = \bar{Y}^{-1}\tilde{s}^+$, $v = \bar{X}\tilde{v}$, $u = \bar{Y}\tilde{u}$, $\pi = \tilde{\pi}$, $\tau = \tilde{\tau}$, $\lambda_j = \tilde{\lambda}_j$, $j \in E$, $d_j = \tilde{d}_j$, $j \in E$ and $b_j = \tilde{b}_j$, $j \in E$, then we will have the following problem

$$
\begin{aligned}
\text{Min} \quad & z_0 = \|s^- - P\pi\|_1 + \|s^+ - Q\tau\|_1 \\
\text{s.t. :} \quad & \\
& \sum_{j\in E} \lambda_j \left(\bar{X}^{-1} X_j\right) = \left(\bar{X}^{-1} X_0\right) - \left(s^- - P\pi\right) && (20.1)\\
& \sum_{j\in E} \lambda_j \left(\bar{Y}^{-1} Y_j\right) = \left(\bar{Y}^{-1} Y_0\right) + \left(s^+ - Q\tau\right) && (20.2)\\
& -v'\left(\bar{X}^{-1} X_j\right) + u'\left(\bar{Y}^{-1} Y_j\right) + d_j = 0 \qquad j \in E && (20.3)\\
& P'v \le 0_{2p} && (20.4)\\
& Q'u \le 0_{2q} && (20.5)\\
& v \ge 1_m && (20.6)\\
& u \ge 1_s && (20.7)\\
& d_j \le M\, b_j \qquad j \in E && (20.8)\\
& \lambda_j \le M\left(1 - b_j\right) \qquad j \in E && (20.9)\\
& \lambda_j, d_j \ge 0, b_j \in \{0,1\} \qquad j \in E && \\
& s^- \ge 0_m, s^+ \ge 0_s, \pi \ge 0_{2p}, \tau \ge 0_{2q} &&
\end{aligned}
\tag{7.20}
$$

Therefore, we can see that (7.18) is equivalent to (7.16) when this latter model is used with the inputs and outputs normalized by their corresponding averages.

7.5 Empirical Illustration

To illustrate the proposed approach, we revisit the example in Ruiz et al. (2015) , which evaluates educational performance of Spanish universities. The study considered 42 public universities (Table 7.2 records the names of such universities), which were evaluated by using the following variables:

Outputs

- GRAD = Graduation rate (y_1): Percentage of students that complete the programme of studies within the planned time, or in one more year, in relation to their entry cohort.
- RET = Retention rate (y_2): It is computed as 100 minus the drop out rate, in order to be treated as an output, i.e., a "the more the better" variable. The drop out rate is the ratio between the number of students of the entry cohort of 2006–07 enrolled for the first time in a subject which do not enrol in their corresponding subjects either in 2007–08 or 2008–09 and the total of students of the entry cohort of 2006–07 (in percent).
- PROG = Progress rate (y_3): Ratio between the number of passed credits[3] corresponding to all the students enrolled in 2008–09 and the total enrolled credits in that academic year (in percent).

[3]Credit is the unit of measurement of the academic load of the subject of a programme.

Inputs

The variables below, which are used as resources, are adjusted according to the number of students in order to take into account the effect of the size of the university. To be precise, the inputs are defined as the ratios between staff (academic and non-academic), expenditure and teaching spaces to the number of students enrolled in 2008–09, measured in terms of equivalent full-time student units (FTStud):

- AStaff (x_1): The ratio between FTAStaff and FTStud, where FTAStaff is the academic staff with full-time equivalence. For example, if AStaff = 0.10 for a

Table 7.2 Universities

COD.	University	COD.	University
UA	U. de ALICANTE	ULPGC	U. de LAS PALMAS DE GRAN CANARIA
UAB	U. AUTÓNOMA DE BARCELONA	UMA	U. de MÁLAGA
UAH	U. de ALCALÁ DE HENARES	UMH	U. MIGUEL HERNÁNDEZ DE ELCHE
UAL	U. de ALMERÍA	UMU	U. de MURCIA
UAM	U. AUTÓNOMA DE MADRID	UOV	U. de OVIEDO
UBA	U. de BARCELONA	UPC	U. POLITÉCNICA DE CATALUÑA
UBU	U. de BURGOS	UPCT	U. POLITÉCNICA DE CARTAGENA
UCA	U. de CÁDIZ	UPF	U. POMPEU FABRA
UCAR	U. CARLOS III DE MADRID	UPM	U. POLITÉCNICA DE MADRID
UCLM	U. de CASTILLA-LA MANCHA	UPN	U. PÚBLICA DE NAVARRA
UCN	U. de CANTABRIA	UPO	U. PABLO DE OLAVIDE
UDG	U. de GIRONA	UPV	U. del PAÍS VASCO
UDL	U. de LLEIDA	UPVA	U. POLITÉCNICA DE VALENCIA
UEX	U. de EXTREMADURA	URI	U. de LA RIOJA
UGR	U. de GRANADA	URV	U. ROVIRA I VIRGILI
UHU	U. de HUELVA	USAL	U. de SALAMANCA
UIB	U. de las ISLAS BALEARES	USC	U. de SANTIAGO DE COMPOSTELA
UJA	U. de JAÉN	USE	U. de SEVILLA
UJCS	U. JAUME I DE CASTELLÓN	UVEG	U. de VALENCIA (ESTUDI GENERAL)
ULC	U. de LA CORUÑA	UVI	U. de VIGO
ULE	U. de LEÓN	UZA	U. de ZARAGOZA

given university, this can be interpreted by saying that there are 10 teachers for each 100 students.

- NAStaff (x_2): The ratio between the total number of administrative and technical support personnel and FTStud.
- EXPEND (x_3): The ratio between expenditure (in euros) and FTStud. Expenditure exactly accounts for staff expenditure, expenditure on goods and services, financial expenditure and current transfers. EXPEND therefore expresses the current expenditure of a given university per student, and reflects the budgetary effort made by the universities in the delivery of their teaching practices. This indicator is traditionally used for comparing institutions, in particular in studies dealing with teaching quality.
- SPAC (x_4): The ratio between the total space in m^2 (corresponding to classrooms, labs and other teaching spaces) and FTStud.

The analysis carried out in Ruiz et al. (2015) incorporated the preferences of experts regarding the relative importance of the variables listed above through the addition of the following AR-I restrictions to the model

$$\begin{aligned}
&2.01 \le \frac{v_{AStaff}}{v_{NAStaff}} \le 10.96 \quad 0.17 \le \frac{u_{RET}}{u_{PROG}} \le 4.48\\
&0.5 \le \frac{v_{AStaff}}{v_{EXPEND}} \le 8.47 \quad 0.09 \le \frac{u_{RET}}{u_{GRAD}} \le 12.48\\
&1.76 \le \frac{v_{AStaff}}{v_{SPAC}} \le 11.15 \quad 0.44 \le \frac{u_{PROG}}{u_{GRAD}} \le 6.30\\
&0.16 \le \frac{v_{NAStaff}}{v_{EXPEND}} \le 3.70\\
&0.27 \le \frac{v_{NAStaff}}{v_{SPAC}} \le 4.51\\
&0.81 \le \frac{v_{EXPEND}}{v_{SPAC}} \le 6
\end{aligned} \tag{7.21}$$

where the weight bounds were obtained from the opinions of experts by using Analytic Hierarchy Process (AHP) (Saaty 1980).

Twelve universities were rated as efficient: UA, UAM, UCAR, UCLM, UGR, UJA, UMA, UPF, URV, USE, UVEG and UVI. It should be noted that UAB, UAL, UBA, UEX, UHU, UPO, UPV and USAL were technically efficient universities that become inefficient as the result of incorporating the expert preferences. Now, we use the proposed approach for benchmarking and setting targets. Specifically, we use model (7.16), including the weight restrictions (7.21), with the data normalized by the averages of the corresponding inputs and outputs and assuming VRS, which means that we must add the convexity constraint $\sum_{j\in E}\lambda_j = 1$ to the formulation and replace the constraints (16.3) with $-v'X_j + u'Y_j + u_0 + d_j = 0, j \in E$, where u_0 is a free variable.

For each of the inefficient universities, Table 7.3 reports the values λ_j^* (the intensities) corresponding to the efficient universities that have participated in its evaluation. $\lambda_j^* = 0$ means that the university "j" has not been a referent in the assessment of the university in the corresponding row, while this role gets more relevant as the value of λ_j^* increases. The last row of this table summarizes the

Table 7.3 Inefficient universities and $\lambda_j^{*\prime}s$ of their benchmarks

Ineff. univ	Efficient university											
	UA	UAM	UCAR	UCLM	UGR	UJA	UMA	UPF	URV	USE	UVEG	UVI
UAB			0.7584					0.2416				
UAH		0.2508				0.0242		0.3205	0.4045			
UAL				0.0918		0.8964		0.0118				
UBA			0.0421		0.6347			0.3232				
UBU								0.1248			0.8752	
UCA			0.2094		0.1722	0.4512		0.1672				
UCN						0.0358		0.5056	0.4028		0.0558	
UDG					0.3243			0.6757				
UDL		0.0407						0.8639	0.0953			
UEX				0.2036		0.5598					0.2365	
UHU		0.3291	0.0260		0.5107			0.1341				
UIB	0.4132				0.5742	0.0126						
UJCS		0.1434			0.6221			0.2345				
ULC						0.5892					0.4108	
ULE						0.4907		0.2881			0.2212	
ULPGC						0.2396	0.0839					0.6765
UMH						0.8158			0.1808		0.0034	
UMU	0.5004				0.1709	0.3287						
UOV								0.1247			0.8753	
UPC								0.9169			0.0831	
UPCT								0.6170			0.3830	
UPM					0.0244	0.3620		0.6135				

(continued)

Table 7.3 (continued)

Ineff. univ	Efficient university											
	UA	UAM	UCAR	UCLM	UGR	UJA	UMA	UPF	URV	USE	UVEG	UVI
UPN								0.6985			0.3015	
UPO			0.4322		0.3069	0.1831		0.0778				
UPV						0.1999			0.1951		0.6050	
UPVA		0.4576				0.0876		0.2753	0.1794			
URI		0.3843				0.0513		0.5206	0.0438			
USAL		0.2350			0.1120	0.1994		0.4536				
USC								0.2549			0.7451	
UZA						0.1416		0.7634			0.0950	
# Referent	2	7	5	2	10	18	1	23	7	0	13	1

number of times each efficient university has acted as referent in the assessments of the others. This provides us with an insight into their role as benchmarks.

UPF and UJA and, to a lesser extent, UVEG and UGR are the universities that have played a more relevant role as benchmarks in the assessment of the inefficient universities. We can see that they have acted as referents in the assessments of 23, 18, 13 and 10 respectively, of the 30 inefficient universities. By contrast, USE played no role as a benchmark, UMA and UVI were referents for only one university and UA and UCLM only for two. UPF is a university with a high level of availability of resources which has also achieved the highest levels in the rates (especially in PROG and GRAD). The results in Table 7.3 show that it has been an important benchmark for the universities with more resources. UJA, UVEG and

Table 7.4 Actual data and targets

Ineff. univ.		Inputs				Outputs		
		AStaff	NAStaff	EXPEND	SPAC	GRAD	RET	PROG
UA	Data	0.0906	0.0631	6965.57	2.0826	28.05	87.11	59.53
UAB	Data	0.1127	0.1070	11,774.05	2.2500	38.58	71.59	73.17
	Targets	0.1046	0.0541	8926.83	2.7956	38.58	75.96	77.56
UAH	Data	0.1099	0.0653	10,104.20	5.1917	38.98	75.19	63.71
	Targets	0.0991	0.0653	10,104.20	5.1917	50.96	83.93	75.33
UAL	Data	0.0924	0.0578	8105.93	3.3098	32.57	87.98	62.67
	Targets	0.0838	0.0409	6708.42	4.9723	38.00	86.83	63.13
UAM	Data	0.1021	0.0471	8926.92	3.6299	43.74	81.00	70.31
UBA	Data	0.1095	0.0667	10,203.00	2.6477	34.70	83.70	69.91
	Targets	0.0912	0.0667	8648.12	3.2034	38.14	83.06	69.91
UBU	Data	0.1110	0.0599	7781.98	7.2993	31.19	79.31	61.81
	Targets	0.0840	0.0533	7781.98	7.2960	45.38	82.90	66.58
UCA	Data	0.0973	0.0513	7798.51	3.8731	31.05	82.00	66.64
	Targets	0.0912	0.0513	7798.51	3.8731	38.03	82.67	68.70
UCAR	Data	0.1037	0.0412	8095.70	2.0956	31.98	73.26	75.65
UCLM	Data	0.0906	0.0544	8135.32	6.0784	45.42	92.81	68.66
UCN	Data	0.1228	0.0740	10,463.68	5.6765	36.65	82.47	65.29
	Targets	0.0987	0.0740	10,463.68	5.6765	53.73	84.71	77.34
UDG	Data	0.1374	0.0858	11,068.05	4.1409	46.38	74.18	71.96
	Targets	0.0994	0.0815	10,134.45	4.1409	49.07	83.98	76.75
UDL	Data	0.1387	0.0889	12,664.43	5.0656	37.00	80.00	72.23
	Targets	0.1060	0.0889	11,275.33	5.0656	57.75	84.38	81.97
UEX	Data	0.0936	0.0480	6189.66	6.7653	39.00	81.00	65.17
	Targets	0.0839	0.0440	7009.00	5.7613	40.16	86.74	64.04
UGR	Data	0.0819	0.0542	7214.39	2.3656	27.77	83.00	62.59
UHU	Data	0.0999	0.0570	8296.78	2.6845	37.47	81.00	63.54

(continued)

Table 7.4 (continued)

Ineff. univ.		Inputs				Outputs		
		AStaff	NAStaff	EXPEND	SPAC	GRAD	RET	PROG
		0.0926	0.0570	8380.68	3.1272	37.37	82.28	68.28
UIB	Data	0.0985	0.0582	7295.94	2.6797	28.00	84.00	61.42
	Targets	0.0855	0.0577	7102.60	2.2800	28.00	84.74	61.32
UJA	Data	0.0828	0.0388	6498.81	4.8588	36.87	86.25	62.30
UJCS	Data	0.1035	0.0627	8984.53	2.7051	32.17	86.61	59.51
	Targets	0.0909	0.0627	8473.41	3.1631	37.45	83.05	68.61
ULC	Data	0.0844	0.0481	6572.31	5.9949	31.68	86.21	57.54
	Targets	0.0819	0.0423	6806.08	5.9949	39.55	84.78	63.06
ULE	Data	0.0912	0.0583	8115.25	5.5092	38.05	81.78	65.74
	Targets	0.0895	0.0568	8115.25	5.5092	44.78	84.94	68.83
ULPGC	Data	0.0889	0.0500	7705.89	3.6760	31.95	80.64	56.00
	Targets	0.0829	0.0435	7009.50	3.6760	31.95	83.61	57.61
UMA	Data	0.0789	0.0533	7178.51	2.8268	20.32	80.43	57.26
UMH	Data	0.0864	0.0416	7558.63	5.1354	36.00	79.00	64.19
	Targets	0.0843	0.0417	7119.03	5.1354	39.21	86.05	64.19
UMU	Data	0.0864	0.0540	6854.68	3.0940	26.09	84.43	59.54
	Targets	0.0865	0.0536	6854.68	3.0434	30.90	86.13	60.96
UOV	Data	0.1016	0.0533	8214.85	8.2193	32.10	78.40	58.73
	Targets	0.0840	0.0533	7781.70	7.2962	45.38	82.90	66.58
UPC	Data	0.1393	0.0907	14,071.92	6.3661	16.24	73.67	69.60
	Targets	0.1055	0.0907	11,179.50	5.2118	57.98	84.30	81.93
UPCT	Data	0.1237	0.0965	9893.47	6.5155	19.70	80.97	51.83
	Targets	0.0973	0.0765	9893.47	6.0007	53.21	83.77	76.12
UPF	Data	0.1078	0.0946	11,536.09	4.9930	59.30	84.45	83.54
UPM	Data	0.1099	0.0813	9606.85	4.8802	8.54	86.00	59.40
	Targets	0.0981	0.0734	9606.85	4.8802	50.41	85.07	75.34
UPN	Data	0.1262	0.0804	10,397.93	10.7657	46.04	82.94	72.59
	Targets	0.0996	0.0804	10,242.98	5.7863	54.51	83.92	77.70
UPO	Data	0.0994	0.0489	7693.97	2.9098	17.28	82.15	70.62
	Targets	0.0935	0.0489	7800.33	2.9098	33.71	79.50	69.81
UPV	Data	0.1155	0.0430	10,062.10	6.8204	43.32	82.16	57.34
	Targets	0.0831	0.0471	7618.02	6.8204	43.32	83.88	65.46
UPVA	Data	0.1101	0.0609	9609.94	4.5986	27.08	84.26	60.97
	Targets	0.1000	0.0609	9609.95	4.5986	48.49	83.16	73.68
URI	Data	0.1139	0.0718	10,203.87	4.5211	43.00	71.36	63.14
	Targets	0.1036	0.0718	10,203.87	4.5211	51.75	83.25	76.89
URV	Data	0.0913	0.0548	9915.63	6.3372	49.68	85.19	72.72
USAL	Data	0.1122	0.0678	9039.19	4.3517	47.64	74.05	68.02
	Targets	0.0986	0.0678	9434.49	4.3517	47.64	83.84	73.85

(continued)

Table 7.4 (continued)

Ineff. univ.		Inputs				Outputs		
		AStaff	NAStaff	EXPEND	SPAC	GRAD	RET	PROG
USC	Data	0.0922	0.0594	9011.39	9.4024	31.45	74.00	64.42
	Targets	0.0875	0.0594	8340.17	6.9536	47.45	83.13	69.10
USE	Data	0.0843	0.0493	7041.43	2.4780	22.56	80.47	60.46
UVEG	Data	0.0806	0.0474	7246.77	7.6244	43.40	82.68	64.16
UVI	Data	0.0835	0.0439	7169.41	3.3624	31.65	83.07	55.99
UZA	Data	0.1401	0.0873	10,415.46	5.2239	35.45	83.37	68.46
	Targets	0.1016	0.0822	10,415.45	5.2239	54.61	84.54	78.69

UGR have lower rates than UPF, but have used much less resources, so they have been important referents for many of the remaining inefficient universities.

For each university, Table 7.4 records its actual data together with the efficient targets for each of the inputs and outputs (for efficient universities only actual data are reported). For some universities, we can see that the targets are very close to their actual data: see the cases of UCA, UHU, UIB, ULE, UMH and UMU. This means that the educational performance of these universities is close to target efficiency. By contrast, the large differences between data and targets for UPC, UPCT and UPM show that these are the most inefficient universities.

Looking at Table 7.4, a general conclusion that can be drawn is that the universities have an important source of inefficiency in AStaff: all of them should, to some extent, reduce the number of teachers per student in order to achieve target efficiency (except the efficient universities). In UBU, UDG, UDL, UPC, UPCT, UPN, UPV and UZA this input should be reduced by more than 20%. As for the outputs, many of the inefficient universities perform weakly regarding the rate GRAD. In universities like UBU, UCN, UDL, UOV, UPO, UPVA, USC and UZA there is room for improvement of more than 40% for this variable. In the case of the polytechnic universities the situation is still more worrying: engineering students in Spain frequently need some more years than those planned to complete their degree, so the rates GRAD (and also PROG) are very low in the universities that are purely polytechnic. In Table 7.4 we can see that the rates GRAD of UPM, UPC and UPCT are, respectively, 8.54% (the minimum across all universities), 16.24 and 19.70% (these figures are much lower than those of other polytechnics outside of Spain). However, their corresponding targets are quite a lot larger, being 50.41, 57.98 and 53.21% respectively. Obviously, the targets that model (16) provides for them are probably unrealistic, perhaps unachievable in the short term if we take into account their starting point. In any case, these results show some aspects of their performance that might need substantial improvements.

Finally, it is worth highlighting that, although there are many similarities between the results provided by (7.16) and those obtained in Ruiz et al. (2015), there are also some differences. For example, UGR seems to play a lesser role as benchmark in the analysis carried out with (7.16), while UAM participates now

more actively in the benchmarking of other universities. This is a consequence of considering in (7.16) the duality relations determined by the weight restrictions. As a result, the targets set by (7.16) are in some cases more demanding (see, in particular, the case of UAL), though the targets provided by both approaches coincide in 16 out of the 30 inefficient universities.

7.6 Conclusions

In management, organizations use benchmarking for the evaluation of their processes in comparison to best practices of others within a peer group of firms in an industry or sector. In the best practice benchmarking process the identification of the best firms enables the setting of targets, which allows these organizations to learn from others and develop plans for improving some aspects of their own performance.

DEA has proven to be a useful tool for the benchmarking. This chapter has dealt specifically with the DEA models with weight restrictions. We claim that the standard restricted DEA models are particularly useful for the benchmarking when weight restrictions are used to incorporate technological judgments into the analysis, in particular information regarding production trade-offs. In those cases, restricted models can be seen as a mean to extend the production possibility set so that targets outside the original one can be considered. Nevertheless, the use of radial models, which do not ensure efficient targets (in the Pareto sense), and the fact that the slacks are maximized (instead of minimized) in the second stage processes that have been proposed, can be seen as a weakness of that approach. As a future research, the formulation of new models that address those issues could be investigated.

In case restricted models are regarded only as a way of incorporating value judgments or preferences, then there may be no reason to argue that targets outside the PPS can be attained. There is therefore a need of models that ensure targets within the PPS appropriately, and this chapter has made a contribution to meeting such need. The models developed find the closest efficient (in the Pareto sense) targets which lie within the PPS. Thus, if weight restrictions reflect preferences or value judgments, the approach proposed allows us to identify best practices that are not only technically achievable (with less effort) but also desirable in the light of prior knowledge and expert opinion.

References

Adler N, Liebert V, Yazhemsky E (2013) Benchmarking airports from a managerial perspective. Omega 41(2):442–458

Allen R, Athanassopoulos A, Dyson RG, Thanassoulis E (1997) Weights restrictions and value judgments in data envelopment analysis: evolution development and future directions. Ann Oper Res 73:13–34

Aparicio J, Pastor JT (2014) Closest targets and strong monotonicity on the strongly efficient frontier in DEA. Omega (United Kingdom) 44:51–57

Aparicio J, Ruiz JL, Sirvent I (2007) Closest targets and minimum distance to the Pareto-efficient frontier in DEA. J Prod Anal 28(3):209–218

Atici KK, Podinovski VV (2015) Using data envelopment analysis for the assessment of technical efficiency of units with different specialisations: an application to agriculture. Omega 54:72–83

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30:1078–1092

Beale EML, Tomlin JA (1970) Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables. In: Lawrence J (ed). Proceedings of the fifth international conference on operational research. Tavistock Publications, London, pp 447–454

Beasley JE (1990) Comparing university departments. Omega 18(2):171–183

Charnes A, Cooper WW (1962) Programming with linear fractional functionals. Naval Res. Logistics Q 9(3–4):181–186

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2(6):429–444

Charnes A, Cooper WW, Thrall RM (1986) Classifying and characterizing inefficiencies in data envelopment analysis. Oper Res Lett 5(3):105–110

Charnes A, Cooper WW, Golany B, Seiford L, Stutz J (1985) Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical productions functions. J Econometrics 30 (1–2):91–107

Charnes A, Cooper WW, Golany B, Seiford L, Stutz J (1987) A dimensionless efficiency measure for departures from pareto optimality. Research report CCS 480, center for cybernetic studies. The University of Texas, Austin

Charnes A, Cooper WW, Huang ZM, Sun DB (1990) Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks. J Econometrics 46(1):73–91

Charnes A, Cooper WW, Lewin AY, Seiford LW (1994) Data envelopment analysis: theory, methodology and applications. Kluwer Academic Publishers, Germany

Cook WD, Tone K, Zhu J (2014) Data envelopment analysis: prior to choosing a model. Omega 44:1–4

Cooper WW, Pastor JT, Borrás F, Aparicio J, Pastor D (2011a) BAM: a bounded adjusted measure of efficiency for use with bounded additive models. J Prod Anal 35(2):85–94

Cooper WW, Park KS, Pastor JT (1999) RAM: a range adjusted measure of inefficiency for use with additive models, and relations to other models and measures in DEA. J Prod Anal 11(1), 5–42

Cooper WW, Seiford LM, Tone K (2007) Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software, 2nd edn. Springer Science & Business Publishers, New York

Cooper WW, Ruiz JL, Sirvent I (2009) Selecting non-zero weights to evaluate effectiveness of basketball players with DEA. Eur J Oper Res 195(2):563–574

Cooper WW, Ruiz JL, Sirvent I (2011b) Choices and uses of DEA weights. In: Cooper WW, Seiford LW, Zhu J (eds) Handbook on data envelopment analysis. Springer, Berlin, pp 93–126

Dai X, Kuosmanen T (2014) Best-practice benchmarking using clustering methods: application to energy regulation. Omega 42(1):179–188

Fukuyama H, Maeda Y, Sekitani K, Shi J (2014) Input–output substitutability and strongly monotonic p-norm least distance DEA measures. Eur J Oper Res 237(3):997–1007

Hung SW, Lu WM, Wang TP (2010) Benchmarking the operating efficiency of Asia container ports. Eur J Oper Res 203(3):706–713

Lins MPE, Sollero MKV, Calôba GM, da Silva ACM (2007) Integrating the regulatory and utility firm perspectives, when measuring the efficiency of electricity distribution. Eur J Oper Res 181 (3):1413–1424

Lovell CAK, Pastor JT (1995) Units invariant and translation invariant DEA models. Oper Res Lett 18(3):147–151

Podinovski VV (2004) Production trade-offs and weight restrictions in data envelopment analysis. J Oper Res Soc 55(12):1311–1322

Podinovski VV (2007a) Improving data envelopment analysis by the use of production trade-offs. J Oper Res Soc 58(10):1261–1270

Podinovski VV (2007b) Computation of efficient targets in DEA models with production trade-offs and weight restrictions. Eur J Oper Res 181(2):586–591

Podinovski VV (2015) DEA models with production trade-offs and weight restrictions. In: Zhu J (ed.) Data envelopment analysis: a handbook of models and methods. Springer, Berlin, pp 105–144

Portela MCAS, Borges PC, Thanassoulis E (2003) Finding closest targets in non-oriented DEA models: the case of convex and non-convex technologies. J Prod Anal 19(2–3):251–269

Portela MCAS, Camanho AS, Borges D (2012) Performance assessment of secondary schools: the snapshot of a country taken by DEA. J Oper Res Soc 63(8):1098–1115

Ruiz JL, Segura JV, Sirvent I (2015) Benchmarking and target setting with expert preferences: an application to the evaluation of educational performance of Spanish universities. Eur J Oper Res 242(2):594–605

Ruiz JL, Sirvent I (2016) Common benchmarking and ranking of units with DEA. Omega 65:1–9

Saaty TL (1980) The analytic hierarchy process. McGraw-Hill International Book Company, USA

Thanassoulis E, Portela MCAS, Allen R (2004) Incorporating value judgments in DEA. In: Cooper WW, Seiford LW, Zhu J (eds) Handbook on data envelopment analysis. Kluwer Academic Publishers, Boston

Thanassoulis E, Portela MCAS, Despić O (2008) Data envelopment analysis: the mathematical programming approach to efficiency analysis. In: Harold, OF, Knox Lovell CA, Schmidt SS (eds.), The measurement of productive efficiency and productivity growth. Oxford University Press, New York, NY (US)

Thompson RG, Singleton FD, Thrall RM, Smith BA (1986) Comparative site evaluations for locating a high-energy physics lab in Texas. Interfaces 16(6):35–49

Thrall RM (1996) Duality, classification and slacks in DEA. Ann Oper Res 66(2):109–138

Thrall RM (2000) Measures in DEA with an Application to the Malmquist Index. J Prod Anal 13 (2):125–137

Tone K (2010) Variations on the theme of slacks-based measure of efficiency in DEA. Eur J Oper Res 200(3):901–907

Wong Y-HB, Beasley JE (1990) Restricting weight flexibility in data envelopment analysis. J Oper Res Soc 41(9):829–835

Zanella A, Camanho AS, Dias TG (2013) Benchmarking countries environmental performance. J Oper Res Soc 64(3):426–438

Chapter 8
Endogenous Common Weights as a Collusive Instrument in Frontier-Based Regulation

Per J. Agrell and Peter Bogetoft

Abstract Non-parametric efficiency analysis, such as Data Envelopment Analysis (DEA) relies so far on endogenous local or exogenous general weights, based on revealed preferences or market prices. However, as DEA is gaining popularity in regulation and normative budgeting, the strategic interest of the evaluated industry calls for attention. We offer endogenous general prices based on a reformulation of DEA where the units collectively propose the set of weights that maximize their efficiency. Thus, the sector-wide efficiency is then a result of compromising the scores of more specialized smaller units, which also gives a more stable set of weights. The potential application could be to precipitate collective bargaining on cost efficiency under regulation with asymmetric information on relative prices and costs. The models are applied to paneldata from 285 Danish district heating plants, where the open evaluation of multiple non-priced outputs is relevant. The results show that sector wide weighting schemes favor input/output combinations that are less variable than would individual units.

Keywords DEA · Efficiency · Weights

JEL Classification D24

P.J. Agrell (✉)
Louvain School of Management and CORE, Universite catholique de Louvain,
1348 Louvain-la-Neuve, Belgium
e-mail: per.agrell@uclouvain.be

P. Bogetoft
Department of Economics, Copenhagen Business School CBS,
Porcelaenshaven 16 A, 2000 Frederiksberg, Denmark
e-mail: pb.eco@cbs.dk

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_8

8.1 Introduction

Weighting of resources consumed and outputs rendered is inherent in any performance evaluation technique that results in a set of measures that is of lower dimensionality than the original production space. The methodology for determining the relative prices is one of the pivotal challenges in performance evaluation. Whereas market prices may be observed or elicited in certain circumstances, they may not necessarily reflect the social welfare effects due to externalities and horizon problems. Technical valuations or specifications may postulate prices for a given technology, but this may be doubtful in regulatory contexts. Non-parametric frontier approaches such as the Data Envelopment Analysis (DEA) by Charnes et al. (1978, 1979) address this issue by allocating sets of *individual endogenous weights* that put the individual unit in the best possible light. In this manner DEA provides the evaluator with a conservative performance estimate that is valid for a range of preference functions. Under a convex frontier specification, the analysis explicitly provides the evaluator with dual information that later may be used to refine the preference model of the evaluator by inserting additional constraints. In an open retrospective evaluation, where the modelling rests entirely at the discretion of the analyst or collectively of the units, such an approach may support organizational learning and development. Unrestricted weights are relevant in the determination of technical efficiency, i.e. the general ability to produce many outputs using few inputs.

Recently, however, DEA has gained a widespread use also in more normative contexts, such as industry and sector evaluations aiming at informing forward-looking decisions in regulation (cf. Agrell et al. 2005), budgeting or incentive management (cf. Agrell et al. 2002). For surveys of applications in the domains of utilities' regulation, see Jamasb and Pollitt (2000), for a full survey cf. Seiford and Thrall (1990). This change of perspective implies higher demands to analyze the strategic behavior of the units, as well as the methodological consistency of the evaluation. A performance measure, albeit conservative, that counter-intuitively discourages relevant economic actions will inevitably lead to dysfunctional behavior. On the other hand, an overly cautious approach using individual dual prices comes at a social cost in terms of the discriminatory capacity of the method.

The use of individual weights is also troublesome from an allocative point of view. If, for example, two units—to put their performance in its best possible light—stipulate the value of labor to capital as being (1:10) and (10:1) respectively, there is clearly a social loss from the allocation of capital and labor. In the DEA literature, allocative efficiency is studied with given market prices. In such cases, allocative efficiency can be evaluated along with technical efficiency to give, for example, the cost efficiency of units. The latter is then an example of *common exogenous weights*. In the absence of unanimous market information that can provide these weights, the DEA approach has usually been to restrain the analysis

to individual endogenous weights, alternatively supplemented with partial price information.

The aim of this paper is the develop a set of *common endogenous weights* that puts the evaluated industry in the best possible light. By using only *ex post* production data, we may derive collective evaluations that are applied across the sample. This will enable us to make cost effectiveness analysis even in cases where relevant market prices do not exist and no other preference information is available. In a normative context, this corresponds to a conservative, yet intra-industry consistent estimate of performance that preempts collective and individual complaints on its validity. Whenever our endogenous common weights are not assessed, the evaluator runs the risk that the evaluated units collectively assert these relative prices and then internally redistributes the allocated incentives among the units. Since the common weights maximize the collective incentive, it also opens for strategic behavior on behalf of the units.

The contribution of the paper is twofold.

First, it extends the methodological discussion on preference modelling in DEA with a treatment of a class of endogenous *and* collective evaluations. A particular application of our approach is in the evaluation of non-balanced activities, where individual weights would have been zerovalued. The common weights here express a comprehensive assessment on these activities, a weighted average of the social benefits.

Second, it addressed a relevant issue in the normative application of DEA in e.g. bargaining or regulation. The suggested approach may be directly used in negotiations with associations that represent the collective of evaluated units.

The outline of this paper is as follows. Section 8.2 presents the traditional approach and derives the individual endogenous weights. Our model is presented in Sect. 8.3, along with some properties and interpretations. An extensive illustration using regulatory panel data from the energy sector is given in Sect. 8.4. The paper is closed with some conclusions in Sect. 8.5.

8.2 Individual Endogenous Weights

In the following we address a traditional setting of evaluated DMUs $j = 1, \ldots, n$, transforming a vector of inputs $x^j \in \mathbb{R}^r_+$ $j = 1, \ldots, n$ to a vector of outputs $y^j \in \mathbb{R}^s_+$ $j = 1, \ldots, n$. Let (X, Y) be the set of observed input-output combinations. The task is to determine a set of input prices $u \in \mathbb{R}^r_+$ and output prices $v \in \mathbb{R}^s_+$ such as to maximize the productivity measure for the unit under evaluation. To avoid degenerated cases, assume that x^j and y^j each contain at least one positive element for all $j = 1, \ldots, n$.

Assuming constant returns to scale as in Charnes et al. (1978), we arrive at the classical "dual" CRS model 8.1:

$$\begin{array}{lll} \max_{u^i,v^i} & \frac{v^i y^i}{u^i x^i} & \\ st & \frac{v^i y^j}{u^i x^j} \leq 1 & \forall j = 1,\ldots,n \\ & u^i \in \mathbb{R}^r_+ \quad v^i \in \mathbb{R}^s_+ & \end{array} \tag{8.1}$$

The optimal solution 8.1 gives a set of individual endogenous weights (u^i, v^i), these are the weights putting DMU i in the best possible light.

In an economic context, the program is equivalent to the maximization of a net profit $v^i y^i - u^i x^i$, given a set of normalized input prices $(u^i x^i = 1)$ and subject to the condition that all observed units run with nonpositive profits $v^i y^j - u^i x^j \leq 0$. See also Pastor et al. (2012) for the development of this equivalence.

$$\begin{array}{lllll} \max_{u^i,v^i} & v^i y^i & & & \\ st & u^i x^i & = & 1 & \\ & v^i y^j - u^i x^j & \leq & 0 & \forall j = 1,\ldots,n \\ & u^i \in \mathbb{R}^r_+ & & v^i \in \mathbb{R}^s_+ & \end{array} \tag{8.1*}$$

Technically, therefore, the weights define a hyperplane that dominate all observed input-output combinations and minimizes the potential improvement in profit by DMU i by doing as well as the best DMUs. Hence, a unit that exhibits a net profit lower than 0 is dominated by some more productive units. An equivalent interpretation of net profit maximization is also found in Pastor et al. (2012).

The dual program 8.1* above is equivalent to the primal program 8.2 below for the decision variables $\left(\theta^i, \lambda\right)$, where θ^i is the radial distance measure for DMU i and λ the convex weights on (X, Y) that dominate (x^i, y^i).

$$\begin{array}{lll} \min_{\theta^i,\lambda} & \theta^i & \\ st & \theta^i x^i \geq & \sum_{j=1}^{n} \lambda_j x^j \\ & y^i \leq & \sum_{j=1}^{n} \lambda_j y^j \\ & \lambda \in \mathbb{R}^n_+ & \end{array} \tag{8.2}$$

The primal description of the production possibility set gives a minimal convex hull that contains all observed units under constant returns to scale. Analogous formulations may also be made under various scale assumptions and distance measures, cf. the cone-ratio approach in Charnes et al. (1989).

It is well known from empirical studies that units under evaluation will claim very diverse prices, since each unit is emphasizing its comparative advantages. This has two implications.

Firstly, in some cases, it is therefore useful (or economically asked for) to introduce an exogenous set of price restrictions. The imposition of restrictions on

the DMU specific prices in DEA models was first proposed by Thompson et al. (1986) as 'assurance regions' (AR) for upper and lower limits on u and v. The AR models developed rapidly by several authors to reflect partial price information, preference information or other subjective information about the relative importance of the inputs and outputs, cf. e.g. Golany (1988), Dyson and Thanassoulis (1988), Wong and Beasley (1990), Roll et al. (1991), Ali et al. (1991), and Halme et al. (1999). For a general discussion of the use of weight restrictions in DEA, see Pedraja-Chaparro et al. (1997). This is easily done by adding constraints on u^i and v^i in the formulation 8.1*. Introducing e.g. $v_h^i \geq v_k^i$ could reflect that output h is at least as important or valuable as output k. More generally, it is useful and straightforward to introduce such information by requiring $u^i \in U^i$ and $v^i \in V^i$, where U^i and V^i are convex polyhedral sets in the strictly positive orthants, $U^i \subset \mathbb{R}^r_{++}$ and $V^i \subset \mathbb{R}^s_{++}$. The resulting version of 8.1* in this case remains a simple linear programming problem.

However, this approach of using exogenous information was challenged by Podinovisky (2004a, b, 2005) arguing that including preference information to derive weight restrictions may lead to inconsistencies and problems in the interpretation of efficiency.

Secondly, it motivates the search for industry-wide prices, which can be expected to be less extreme than the individual prices. However, as we shall see in the numerical illustration, this depends on the underlying technology.

8.3 Common Endogenous Weights

We now revisit the classical CRS model to determine a common set of weights (u, v) for all units, so that the overall efficiency of the set of units is maximized. Consider the following program $P3$:

$$\begin{array}{lllll} \max_{u,v} & \frac{v\sum_{i=1}^{n} y^i}{u\sum_{i=1}^{n} x^i} & & & \\ st & \frac{vy^j}{ux^j} & \leq & 1 & \forall j = 1, \ldots, n \\ & \frac{v\sum_{i=1}^{n} y^i}{u\sum_{i=1}^{n} x^i} & \leq & 1 & \\ & u \in \mathbb{R}^r_+ & & v \in \mathbb{R}^s_+ & \end{array} \tag{8.3}$$

where the objective function expresses the aggregate productivity, the first constraint the individual productivity normalization as in CRS and the second constraint the collective normalization. The interpretation is straightforward. We seek the prices that make the joint production plan look as attractive as possible, subject to the usual normalization constraints that no observed production, joint or individual, can have a benefit-cost ratio exceeding 1. The results in model P3 are mathematically equivalent to the centralized resource allocation case for CRS in

Lozano and Villa (2004). However, we formulate the models in our notation for consistency.

Remark 1 The second constraint $\frac{v\sum_{i=1}^{n} y^i}{u\sum_{i=1}^{n} x^i} \leq 1$ is redundant.

Proof $vy^j/ux^j \leq 1$ implies $vy^j - ux^j \leq 0$ for all j, which implies $\sum_{j=1}^{n}(vy^j - ux^j) \leq 0$. ∎

In turn this implies $\sum_{j=1}^{n} vy^j / \sum_{j=1}^{n} ux^j \leq 1$.

Remark 2 One of the n first constraints will always be binding.

Proof Assume that (u^*, v^*) is an optimal solution to the program and that none of the n first constraints are binding, i.e. $v^*y^j/u^*x^j < 1$ for all j. This implies $v^*y^j - u^*x^j < 0$ for all j and we may therefore increase all elements of v^* marginally without violating the constraints. This would increase $\sum_{j=1}^{n} v^*y^j / \sum_{j=1}^{n} u^*x^j$ and it thus contradicts the optimality of (u^*, v^*). ∎

An alternative interpretation of the objective function is stated without its (trivial) proof.

Remark 3 The objective function can be rewritten as a weighted average of the usual benefit cost ratios: $\sum_{i=1}^{n}\left(\frac{ux^i}{\sum_{j=1}^{n} ux^j}\frac{vy^i}{ux^i}\right)$

The solution to program 8.3 above is not unique. The normalization is (as well as in the original CRS model) arbitrary and any inflation or deflation of all prices with the same factor would not affect the solution.

The problem 8.3 can also be reformulated as a game theoretic problem 8.4, where the industry picks prices to maximize the value added and the regulator selects the benchmark ratio, as the most promising from the set of individual processes:

$$\begin{array}{ll} max_{u,v} & \frac{v\sum_{i=1}^{n} y^i}{u\sum_{i=1}^{n} x^i} / \max_j\left\{\frac{vy^j}{ux^j}\right\} \\ st & u \in \mathbb{R}^r_+, \quad v \in \mathbb{R}^s_+ \end{array} \tag{8.4}$$

However, a more conventional primal reformulation of $P3$ is given below as 8.5, which is equivalent to the superefficiency (Andersen and Petersen 1993) evaluation of an aggregate unit $\left(\sum_{i=1}^{n} x^i, \sum_{i=1}^{n} y^i\right)$. The equivalence between $P3$ and $P5$ is proved in Proposition 1 below.

$$\begin{array}{lllll} \min_{\theta,\lambda} & \theta & & & \\ st & \theta\left(\sum_{i=1}^{n} x^i\right) & \geq & \sum_{j=1}^{n} \lambda^j x^j & \\ & \sum_{i=1}^{n} y^i & \leq & \sum_{j=1}^{n} \lambda^j y^j & \\ & \lambda \in \mathbb{R}^n_+ & & & \end{array} \tag{8.5}$$

Here, the interpretation falls out immediately from the formulation as the amount of the pooled inputs that could have been saved in the production of the pooled outputs by allocating the production in the best possible way among the available production processes. Assuming proportional weights, it is also the optimal production plan in a centralized planning setting with given subprocesses.

The primal formulation in 8.5 is also related to the measure of overall potential gains from mergers developed in Bogetoft and Wang (2005) and Bogetoft et al. (2003), and to the measures of structural efficiency originally suggested by Farrell (1957) and Førsund and Hjalmarsson (1979).

Proposition 1 *The dual variables associated with the two sets of constraints in will be the optimal weights or prices u and v in 8.3.*

Proof Usual dualization of 8.5 gives the program

$$\begin{array}{lllll} \max_{u,v} & v\left(\sum_{i=1}^{n} y^i\right) & & & \\ st & u\left(\sum_{i=1}^{n} x^i\right) & \leq & 1 & \\ & vy^j - ux^j & \leq & 0 & \forall j = 1,\ldots,n \\ & u \in \mathbb{R}^r_+, v \in \mathbb{R}^s_+ & & & \end{array} \tag{8.6}$$

Without loss of generality, we may require that $u\left(\sum_{i=1}^{n} x^i\right) = 1$ and program *P*6 is thus equivalent to

$$\begin{array}{lllll} \max_{u,v} & \frac{v\sum_{i=1}^{n} y^j}{u\sum_{i=1}^{n} x^i} & & & \\ st & \frac{vy^j}{ux^j} & \leq & 1 & \forall j = 1,\ldots,n \\ & u\sum_{i=1}^{n} x^i & = & 1 & \\ & u \in \mathbb{R}^r_+ & & v \in \mathbb{R}^s_+ & \end{array}$$

Again, this is equivalent to *P*3 due to redundancy of the second constraint in *P*3 (Remark 4) and the possibility to scale all prices. ∎

As can be easily seen from the program 8.6, the common weight problem essentially maximizes the total payment the industry can claim, given the knowledge that the regulator has about the best production practices. Hereby, any Pareto efficient solution for the industry is supported and could potentially be implemented using appropriate sidepayments. This suggests that if the industry is bargaining for incentive payments based on the prices *u*, *v* the units should collectively agree on the common weights. Note also from the game-theoretical formulation above that the bargaining power of the regulator is given by his information about the efficiency of individual processes. This limits the rents the collective of units can claim using extreme weights.

So far we have not made use of, nor assumed the existence of, market prices of the inputs or outputs. The revenue and cost terms calculated are merely used to define the reimbursement scheme. An interesting possibility is that the attempt to optimize incentives under DEA control may be costly when considering the true market prices. This is the case if the reduced incentive cost forces the DMU away from the locally allocatively efficient productions.

8.4 Weight Restrictions in Regulation

Although frontier regulation is popular, especially in European network regulation cf. Agrell and Bogetoft (2016), weight restrictions have to our knowledge been used very rarely: in Norway for electricity distribution system operators in 2008 (cf. Bjorndal et al. 2010) and in the international benchmarking for electricity transmission (e^3GRID 2012), cf. Agrell and Bogetoft (2014). The exogenous common weight restrictions in Norway were abandoned in the subsequent model specification for the DEA yardstick model due to unintended effects on the frontier assessments, similar to those reported in the literature above.

The setting for the international electricity transmission in Agrell and Bogetoft (2014) was different. By default the number of DMUs in the reference set is extremely limited ($n = 22$) and a large part of the benchmarking consists in various standardization and normalization operations on the costs and technical assets reported in the study. The three output variables in the model are described in Table 8.1 below for an average cost (total expenditure) model. The variable *NormGrid* is basically a weighted sum of the relevant transmission assets that the operators provide to the system, *DenseArea* is the total area (in km^2) of city

Table 8.1 Average cost model for electricity transmission in Europe, dependent variable log (TOTEX), OLS and robust OLS, Agrell and Bogetoft (2014)

	OLS	Robust OLS
log(*NormGrid*)	0.554*** (0.096)	0.475*** (0.052)
log(*DenseArea*)	0.117*** (0.011)	0.137*** (0.009)
log(*AngleLineSum*)	0.217** (0.083)	0.284*** (0.040)
Constant	9.233*** (0.516)	9.477*** (0.338)
Observations	102	
Adjusted R^2	0.912	
Residual Std. Error	0.351 (df = 98)	
F Statistic	349.668*** (df = 3; 98)	

Note *p <0.1; **p < 0.05; ***p < 0.01

population density (EUROSTAT definition) and *AngleLineSum* is the weighted linelength of angular towers. The first variable is the principal cost driver, the two others are proxies for the extra operating and capital costs due to high infrastructure density and routing complexity. The technology in electricity transmission was shown to exhibit economies of scale and an non-decreasing returns to scale model (NDRS) was used in DEA for the study. Given the size of the sample and the obvious observation that some operator by default has the highest or lowest values for each variable, an unrestricted frontier would give very low discrimination among the operators. The endogenous dual weights for the extreme operators are heavily biased to the specific outputs for which they dominate. From an engineering-economic viewpoint this is not sensible since e.g. a higher incidence of angular towers cannot increase the capital and operating costs by any proportion for the parts of the grid that are not concerned by these complexities. It was therefore decided to apply weight restrictions in form of cones around the endogenous dual weights for the two supplementary variables corresponding to a confidence interval bands at the 99% level. The final results from the study were used in regulation and their use was endorsed by appeal court rulings in the Netherlands in 2014.

In the actual application, the operators are competing and no data are shared. The task of defining the limits for trade-offs on the frontier is consequently relegated to the consultants. However, the ideal information would of course have been that concerning the exact trade-off ratios for an efficient cost-minimizing firm. In line with the idea in this paper, suppose that there would be agreement among the operators that the trade-off ratios are identical, but not their values. Facing this additional information, the pressure on the regulators would have been high to accept and implement this information in the benchmarking. The use of a common endogenous dual weight set would be the solution. What shape does this information have in a collusive setting? What type of results can be expected? These questions will be answered in the next section where we revisit another application.

8.5 Numerical Illustration

To illustrate the proposed model, we use the paneldata in Agrell and Bogetoft (2004) of district heating plants in Denmark. Define x_1 as the operating expenditure in MDKK, x_2 as the primary fuel input in GJ, y_1 as the heat energy delivered in GJ, y_2 as the electrical energy delivered in GWh, y_3 as the heat capacity utilized in MW, y_4 as the total length of pipelines. The input-oriented model uses the non-discretionary pipeline length as a proxy for customer density. The dataset contains 285 DMUs for 1998/99 and 234 DMUs for 1999/00, each representing the performance of a district heating plant. Real input prices w_1, w_2 are known for each DMU. Some descriptive statistics about the data are presented in Tables 8.2 and 8.3.

The following efficiencies under constant returns to scale are assessed for each yearly dataset: input-oriented technical efficiency $TE(year)$ given by a formulation,

Table 8.2 Descriptive statistics, district heating plants in Denmark 1998/99, Agrell and Bogetoft (2004)

		Mean	Median	Min	Max	Standard.dev.
x_{opex}	kDKK	6319	2301	138	226,039	18,580
$w_{fuel}x_{fuel}$	kDKK	17,998	4633	194	854,646	65,698
x_{fuel}	GJ	314	84	10	11,919	1037
c	kDKK	24,318	7247	625	961,049	78,604
z_{pipes}	km	54	25	0	1597	127
y_{heat}	GJ	259	64	7	10,308	907
y_{elec}	GWh	9	0	0	280	24
y_{cap}	MW	78	21	2	2405	241

Table 8.3 Descriptive statistics, district heating plants in Denmark 1999/00, Agrell and Bogetoft (2004)

		Mean	Median	Min	Max	Standard.dev.
x_{opex}	kDKK	7550	2535	60	268,941	23,724
$w_{fuel}x_{fuel}$	kDKK	21,490	5577	238	932,759	76,421
x_{fuel}	GJ	352	86	10	12,018	1145
c	kDKK	29,040	8119	693	1,068,944	93,260
z_{pipes}	km	63	26	0	1629	147
y_{heat}	GJ	296	64	1	10,658	1050
y_{elec}	GWh	10	0	0	267	25
y_{cap}	MW	87	21	2	2499	273

cost efficiency given local prices $CE(w^{year})$, cost efficiency given average input prices $CE(\overline{w}^{year})$, aggregate cost efficiency $ACE(w^{year})$ and cost efficiency under the new approach $CEC(u^{*,year}, v^{*,year})$ using the formulation. We also introduce the notation $C(w) = wx$ for the total realized cost at valuation w. For clarity, we restate the programs $CE(.)$ and $ACE(.)$ below.

$$CE^i\left(w^i\right) = \min\left\{ w^i \sum_{j=1}^{n} \lambda^j x^j \middle| \sum_{j=1}^{n} \lambda^j x^j \le x^i, \sum_{j=1}^{n} \lambda^j y^j \ge y^i, \lambda \in \mathbb{R}^n_+ \right\}$$

$$ACE^i\left(w^i\right) = \min\left\{ \sum_{j=1}^{n} \left(w^j x^j\right) \lambda^j \middle| \sum_{j=1}^{n} \lambda^j \left(w^j x^j\right) \le w^i x^i, \sum_{j=1}^{n} \lambda^j y^j \ge y^i, \lambda \in \mathbb{R}^n_+ \right\}$$

The aggregate units $x^{98} = \sum_{j=1}^{285} x^{j,98}$ and $x^{99} = \sum_{j=1}^{234} x^{j,99}$ and corresponding output aggregations are included in the calculations, but are inefficient under any assumptions.

The results of the assessments are presented in Table 8.4. As expected, the technical efficiency estimates $TE(.)$ are the highest, well above 0.80 for even this fairly aggregated two-input model. The proposed common weights model in

Table 8.4 Cost efficiency estimates for various model specifications

	n	Mean	r	Min	Standard dev.
$TE(98)$	285	0.827	24	0.610	0.093
$TE(99)$	234	0.814	19	0.330	0.107
$CEC_{98}(u^{98}, v^{98})/C(u^{98})$	285	0.805	11	0.600	0.088
$CEC_{99}(u^{99}, v^{98})/C(u^{99})$	234	0.792	11	0.290	0.105
$CE_{98}(\overline{w}^{98})/C(\overline{w}^{98})$	285	0.745	9	0.440	0.122
$CE_{99}(\overline{w}^{99})/C(\overline{w}^{99})$	234	0.737	11	0.310	0.136
$CE_{98}(w^{98})/C(w^{98})$	285	0.730	9	0.120	0.139
$CE_{99}(w^{99})/C(w^{99})$	234	0.725	12	0.160	0.160
$ACE_{98}(w^{98})/C(w^{98})$	285	0.600	7	0.300	0.137
$ACE_{99}(w^{99})/C(w^{99})$	234	0.614	10	0.260	0.144

Table 8.4 denoted $CEC(u^*, v^*)$ yields in both cases the degenerated dual prices $u^* = \{0, 1\}$ and $v^* = \{1, 0, 0, 0\}$, effectively transforming the efficiency problem into a two-dimensional issue of heat losses. As the heat losses are limited by the thermophysical configuration of the network, the overall efficiencies are high. Inputs such as operating expenditure have higher variability towards output, which favors selected DMUs, but lower the bulk of the scores. Irrespective of whether an extremely favorable observation is by skill or luck, the regulator-evaluator is using this variability to gauge all units in DEA. The industry collectively can hedge itself against this bargaining power by de-emphasizing inputs (e.g. operating expenditures) with higher variability towards outputs in favor of the inputs that are naturally bounded by proportionality. This explains the seemingly counterintuitive result. The lowest individual score for 1999, 0.290, is likely due to reporting errors rather than actual heatlosses. The $CE(\overline{w})$ model assesses the cost efficiency under the premises that all units are subject to average fuel prices. As the majority of the units are single sourced heat plants, this assumes change of technology. The resulting scores are lower, around 0.74, as the trade-off between operating expenditure and fuel cost becomes more realistic. The next model $CE(w)$ changes the competition by exposing DMUs to evaluation by local prices, which of course may be lower or higher than the market average. The efficiency level is roughly as with average prices, albeit with some extreme dips for certain technologies. Finally, the $ACE(w)$ model assumes that the markets behind the DMUs may purchase power and heat to average prices from any other units. Here, the integrated efficiency of operating expenditure, fuel purchases and fuel efficiency is estimated. In this case, the resulting efficiencies are low, around 0.60, highlighting the large discrepancies in overall cost efficiency on the district heat market.

An interesting break-down of the results for the five models is made in Table 8.5. The two first columns give the minimal cost estimates for each model, in applicable prices. The four columns to the right tabulate the optimal production profile for each model, and the current production is given in the top row.

Table 8.5 Minimal cost estimates and input requirements

	$C(w)$		OPEX		Fuel	
	1998	1999	1998	1999	1998	1999
x	6856	6804	1795	1770	89	82
$x^{u}(u^*, v^*)$	6047	5637	1846	1489	74	68
$x^{TE}(w)$	5878	5711	1520	1478	76	69
$x^{CE}(\overline{w})$	5425	5247	1096	895	76	71
$x^{CE}(w)$	5390	5217	1016	927	77	71
x^{ACE}	4228	4276	–	–	–	–

As expected from the model formulation, the proposed model minimizes the overall inefficiency, as $C(u) = x_2$. However, to assess the budget value $C(w)$, the average prices $\overline{w}$ are used. Note that this model implies a hefty substitution rate between opex and fuel, reducing the fuel input to an absolute minimum. The technical efficiency model *TE* implies a proportional reduction of both inputs, disregarding the actual substitution rate. The overall budget under local prices w is lower than for the common weights model, the results are marginally different than for average prices $\overline{w}$. The case illustrates thus the difference between the objective to maximize average efficiency (as for *TE*) and to maximize aggregate efficiency. The 64 DMUs that have lower score in the new aggregated model than in the technical efficiency model are primarily units that are comparatively stronger in partial opex-efficiency. In one outlier case, a small technically efficient unit drops to 0.51 in aggregate efficiency. The following two models $CE(\overline{w})$ and $CE(w)$ suggest substantial reductions of opex at comparatively higher levels of fuel than the aggregate model. The detailed outcome of these models give raise to far more revolutionary changes of the organization and technology in the market than the aggregate and technical efficiency models. The overall cost efficiency model $ACE(w)$ evaluated at local budgets, indicates a further 1000 MDKK reduction of controllable costs, which of course presumes complete flexibility in scale and scope of operations.

It is interesting to note that the costs using local prices may well increase when using the production plan generated using common weights incentives, $x^{u}(u^*, v^*)$. This reflects the potential conflict of reducing incentive costs and achieving allocative efficiency.

8.6 Conclusions

In this paper we derive endogenous sector-wide prices for DEA evaluations. This is useful when there are no exogenous general weights available, nor relevant to use local endogenous prices. The resulting model can be interpreted as a game theoretic model, where the industry suggests prices to collectively maximize net revenue or compensation and a principal selects a benchmarking unit to constrain the set of

acceptable prices. This interpretation is specifically valid in the frequent scenarios evoked in applied work where the regulated or evaluated units are 'consulted' in order to derive or validate the specification of the activity model(s) used in the assessment. The common weight approach can then be seen as a focal point for the firms in a cooperative game against the evaluator, providing a basis for the side payments necessary to implement the solution.

We illustrate the model using paneldata from Danish district heating plants. The outcome has several intriguing characteristics, among those the risk reducing strategy of emphasizing input-output dimensions with low variability across the sample. The empirical study also illustrates the potential direct distortion of total allocative efficiency when reacting strategically to collective incentives. Further work intends to explore the cooperative game properties of specification of models for performance assessment. Another avenue for further research is the normative use of common weights for resource allocation and target setting under strategic uncertainity, such as in Hatami-Marbini et al. (2015). The group-wise approach in Cook and Zhu (2007) could be interesting also from a strategic viewpoint if the units may choose their group assignment. Another promising aspect for regulatory applications with small datasets is to explore the results in Thanassoulis and Allen (1998) where the weight restrictions are made equivalent to new artificial observations.[1]

Acknowledgements The authors would like to thank Peter Fristrup, Yves Pochet and an anonymous referee for useful comments and suggestions on a previous version of this paper.

References

Agrell PJ, Bogetoft P (2004) Economic and environmental efficiency of district heating plants. Energy Policy 33(10):1351–1362

Agrell PJ, Bogetoft P (2014) International benchmarking of electricity transmission system operators. In: 11th international conference on the European energy market (EEM14). IEEE, pp 1–5. doi:10.1109/EEM.2014.6861311

Agrell PJ, Bogetoft P (2016) Regulatory benchmarking: Models, analyses and applications. DEA J (forthcoming)

Agrell PJ, Bogetoft P, Tind J (2002) Incentive plans for productive efficiency, innovation and learning. Int J Prod Econ 78:1–11

Agrell PJ, Bogetoft P, Tind J (2005) Dynamic DEA and yardstick regulation in Scandinavian electricity distribution. J Prod Anal 23(2):173–201

Ali AI, Cook WD, Seiford LM (1991) Strict vs. weak ordinal relations for multipliers in data envelopment analysis. Manage Sci 37:733–738

Andersen P, Petersen NC (1993) A procedure for ranking efficient units in data envelopment analysis. Manage Sci 39:1261–1264

Bjorndal E, Bjorndal M, Fange KA (2010) Benchmarking in regulation of electricity networks in Norway: an overview. In: Energy, natural resources and environmental economics. Springer, Berlin, pp 317–342

[1]Forsund (2013) recalls that already the seminal paper by Farrell (1957) draws on artificial units to impose the right curvature for the isoquants.

Bogetoft P, Wang D (2005) Estimating the potential gains from mergers. J Prod Anal 23:145–171
Bogetoft P, Thorsen BJ, Strange N (2003) Efficiency and merger gains in the Danish forestry extension service. For Sci 49(4):585–595
Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444
Charnes A, Cooper WW, Rhodes E (1979) Short communication: measuring the efficiency of decision making units. Eur J Oper Res 3:339
Charnes A, Cooper WW, Wei Q, Huang Z (1989) Cone ratio data envelopment analysis and multi-objective programming. Int J Syst Sci 20:1099–1118
Cook WD, Zhu J (2007) Within-group common weights in DEA: an analysis of power plant efficiency. Eur J Oper Res 178(1):207–216
Dyson RG, Thanassoulis E (1988) Reducing weight flexibility in data envelopment analysis. J Oper Res Soc 6:563–576
Farrell MJ (1957) The measurement of productive efficiency. J Roy Stat Soc 120:253–281
Førsund FR (2013) Weight restrictions in DEA: misplaced emphasis? J Prod Anal 40(3):271–283
Førsund F, Hjalmarsson L (1979) Generalized Farrell measures of efficiency: an application to milk processing in Swedish dairy plants. Econ J 89:294–315
Golany B (1988) A note on including ordinal relations among multipliers in data envelopment analysis. Manage Sci 34:1029–1033
Halme M, Joro T, Korhonen P, Salo S, Wallenius J (1999) A value efficiency approach to incorporating preference information in data envelopment analysis. Manage Sci 45(1):103–115
Hatami-Marbini A, Tavana M, Agrell PJ, Lotfi FH, Beigi ZG (2015) A common-weights DEA model for centralized resource reduction and target setting. Comput Ind Eng 79:195–203
Jamasb T, Pollitt M (2000) Benchmarking and regulation: international electricity experience. Utilities Policy 9(3):107–130
Lozano S, Villa G (2004) Centralized resource allocation using data envelopment analysis. J Prod Anal 22(1–2):143–161
Pastor JT, Lovell CK, Aparicio J (2012) Families of linear efficiency programs based on Debreu's loss function. J Prod Anal 38(2):109–120
Pedraja-Chaparro F, Salinas-Jimenez J, Smith P (1997) On the role of weight restrictions in data envelopment analysis. J Prod Anal 8:215–230
Podinovski VV (2004a) Production trade-offs and weight restrictions in data envelopment analysis. J Oper Res Soc 55(12):1311–1322
Podinovski VV (2004b) Suitability and redundancy of non-homogeneous weight restrictions for measuring the relative efficiency in DEA. Eur J Oper Res 154(2):380–395
Podinovski VV (2005) The explicit role of weight bounds in models of data envelopment analysis. J Oper Res Soc 56(12):1408–1418
Roll Y, Cook WD, Golany B (1991) Controlling factor weights in data envelopment analysis. IIE Trans 23:2–9
Seiford LM, Thrall RM (1990) Recent developments in DEA: the mathematical programming approach to frontier analysis. J Econometrics 46:7–38
Thanassoulis E, Allen R (1998) Simulating weights restrictions in data envelopment analysis. Manage Sci 44:586–594
Thanassoulis E, Dyson RG (1992) Estimating preferred target input-output levels using data envelopment analysis. Eur J Oper Res 56:80–97
Thompson RG, Singleton FD Jr, Thrall RM, Smith BA (1986) Comparative site evaluations for locating a high-energy physics lab in Texas. Interfaces 16(6):35–49
Wong Y-HB, Beasley JE (1990) Restricting weight flexibility in data envelopment analysis. J Oper Res Soc 41:829–835

Chapter 9
A Parameterized Scheme of Metaheuristics to Solve NP-Hard Problems in Data Envelopment Analysis

Juan Aparicio, Martin Gonzalez, Jose J. Lopez-Espin and Jesus T. Pastor

Abstract Data Envelopment Analysis (DEA) is a well-known methodology for estimating technical efficiency from a set of inputs and outputs of Decision Making Units (DMUs). This paper is devoted to computational aspects of DEA models when the determination of the least distance to the Pareto-efficient frontier is the goal. Commonly, these models have been addressed in the literature by applying unsatisfactory techniques, based essentially on combinatorial NP-hard problems. Recently, some heuristics have been introduced to solve these situations. This work improves on previous heuristics for the generation of valid solutions. More valid solutions are generated and with lower execution time. A parameterized scheme of metaheuristics is developed to improve the solutions obtained through heuristics. A hyper-heuristic is used over the parameterized scheme. The hyper-heuristic searches in a space of metaheuristics and generates metaheuristics that provide solutions close to the optimum. The method is competitive versus exact methods, and has a lower execution time.

Keywords Data envelopment analysis · Closest targets · Mathematical programming · Metaheuristics · Parameterized scheme

9.1 Introduction

Data Envelopment Analysis (DEA) is a well-known technique to estimate the level of efficiency of a set of firms or organizations and, in general, a group of Decision Making Units (DMUs). Efficiency evaluation in production has been an important issue for managers as well as an area of interest from a practical and methodological view in operations research and economics. The focus of such assessment is to analyze the technical efficiency of a DMU, which uses several inputs to

J. Aparicio (✉) · M. Gonzalez · J.J. Lopez-Espin · J.T. Pastor
Center of Operations Research, University Miguel Hernandez,
Elche, Alicante, Spain
e-mail: j.aparicio@umh.es

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_9

produce several outputs, by comparing its performance with respect to the boundary of an estimated production possibility set, using to that end a sample of other DMUs operating in a similar technological environment. In production settings where only one output is produced, the feasible production plans are summarized in the notion of production function, which represents the maximum product obtainable from the input combination at the existing state of technical knowledge.

The estimation of production functions from a data sample began in the area of economics with the application of regression analysis and Ordinary Least Squares to estimate a parametrically specified 'average' production function (Cobb and Douglas 1928). Later, Farrell (1957) showed how to estimate an isoquant enveloping all the observations, overcoming the problem of determining an average function that, consequently, does not meet the basic requirements of a production function. Farrell's paper represented an enormous advance in the measurement of production efficiency and the starting point for numerous subsequent approaches in the field of efficiency analysis. Farrell inspired other authors to continue this line of research estimating production functions that envelop all the observations of the sample by either a non-parametric piecewise linear technology or a parametric function. The first possibility was taken up by Charnes et al. (1978), Banker et al. (1984) and others, resulting in the development of DEA, whereas the latter approach was taken up by Aigner and Chu (1968), Aigner et al. (1977) and Meeusen and van den Broeck (1977) and others, subsequently resulting in the development of deterministic and stochastic frontier models.

In contrast to the deterministic and stochastic frontier models, DEA is a non-parametric technique, since it is not necessary to postulate a specific functional form for the production function. Additionally, DEA is based on mathematical programming, mainly Linear Programming (LP), generating polyhedral technologies where the frontier is, therefore, piecewise linear. There are different DEA efficiency measures, depending on the way that the distance from the evaluated DMU to the frontier of the technology wants to be implemented. Indeed, the first years of life of DEA witnessed the introduction of many different technical efficiency measures, such as the Russell input and output measures of technical efficiency and their graph extension (see Färe et al. 1985), the additive model (Charnes et al. 1985), the Range-Adjusted Measure (Cooper et al. 1999) and the Enhanced Russell Graph (Pastor et al. 1999) or Slacks-Based Measure (Tone 2001), to name but a few. The reason for the introduction of many different technical efficiency measures is the piecewise linear nature of the boundary of the production possibility set in DEA. In this setting, Pareto-efficiency (Koopmans 1951) comes into play. In fact, certain DEA measures, as for example the additive model by Charnes et al. (1985), have been introduced for ensuring that the evaluated units were compared exclusively with respect to the set of Pareto-efficient points (the set of non-dominated points of the DEA technology), also known as the strongly efficient frontier. Moreover, each efficiency measure in DEA is calculated by solving a model of mathematical programming. DEA models provide both an efficiency score for each of the assessed DMUs and information

on the targets that have been used in the efficiency assessment in the case of dealing with inefficient DMUs. The targets are the coordinates, in the input-output space, of the efficient projection point on the frontier and thus represent levels of operation of inputs and outputs that would make the corresponding inefficient DMU perform efficiently. Consequently, targets are highly relevant from a managerial point of view.

Most traditional DEA efficiency measures yield targets that are located far from the evaluated unit (see, for example, Aparicio et al. 2007), being too exacting and not easily achievable by DMUs. This drawback has generated an increasing interest of researchers to develop DEA measures of technical efficiency that are capable of yielding more suitable targets. The philosophy behind all these approaches is the application of the Principle of Least Action (PLA), which always seeks the closest efficient targets to the assessed DMU (Aparicio et al. 2014a). Let us now briefly review the main papers on this approach. It seems that all started with Frei and Harker's paper (1999), where the main objective was to determine projection points by minimizing the Euclidean distance to the strongly efficient frontier in DEA. Later, Cherchye and Van Puyenbroeck (2001) defined the deviation between mixes in a space-oriented framework as the angle between the input vector of the assessed DMU and its projection, maximizing the corresponding cosine in order to find the closest targets. Gonzalez and Alvarez (2001) redefined the classical input-oriented Russell efficiency measure (Färe et al. 1985) based on the minimization of the sum of input contractions required to reach the efficient subset of the production frontier. Silva et al. (2003) introduced the notion of 'similarity' as closeness between the values of inputs and/or outputs of the evaluated DMU and those of the obtained projection (the targets), and they consequently suggested finding projection points as similar as possible to the assessed unit. Lozano and Villa (2005) introduced a method that determines a sequence of targets to be achieved in successive steps, which converge on the strongly efficient frontier. Aparicio et al. (2007) determined the closest targets for a set of international airlines by applying a new version of the Enhanced Russell Graph/Slacks-Based Measure, characterizing the Pareto-efficient frontier. More recently, Baek and Lee (2009), Amirteimoori and Kordrostami (2010) and Aparicio and Pastor (2014a) have focused closely on the determination of a weighted Euclidean distance to the strongly efficient frontier and have showed the fulfillment of certain properties. This topic alone, the satisfaction of a set of interesting properties, mainly monotonicity, has motivated the recent publication of several papers in the context of least distance calculation: Pastor and Aparicio (2010), Ando et al. (2012), Aparicio and Pastor (2013), (2014a, b), Fukuyama et al. (2014a, b), (2016).

In general, applying the approach based on the Principle of Least Action is computationally more difficult than obtaining the furthest efficient targets (the classical approach), since the latter are usually associated with the resolution of a standard linear program, something that does not happen with the determination of the least distance to the production frontier. In particular, minimizing the distance from an inefficient DMU to the frontier is equivalent to calculating the distance

from an interior point of the polyhedral technology to the complement of a convex set, which is not a straightforward problem (see Briec 1997). In fact, some published strategies have been based on a multi-stage approach. The first stage consists in determining all the efficient faces of the DEA frontier, while in a second stage the selected measure is computed for each face, reporting, finally, the least distance measure.

As for papers that have studied the computational aspects of DEA models associated with the PLA, we draw attention to Aparicio et al. (2014b), Jahanshahloo et al. (2005, 2007, 2012), Benavente et al. (2014), López-Espín et al. (2014) and Gonzalez et al. (2015). Some of these approaches are based on Mixed Integer Linear Programming or Bilevel Linear Programming; others are derived from algorithms that allow the determination of all the facets of a polyhedron whereas, finally, some others apply genetic algorithms.

The focus of this chapter is to show how it is possible to solve and study the NP-hard problems associated with the approach based on the determination of closest targets resorting to genetic algorithms and heuristics. To implement this objective, we will resort to a particular DEA efficiency measure, known as the Enhanced Russell Graph Measure or Slacks-Based Measure (Pastor et al. 1999; and Tone 2001). This strategy will allow us to show specific algorithms and results related to them. Indeed, this new measure, which applies the PLA, has already been analyzed in some recent papers. In particular, in Aparicio et al. (2007), a new version of the Enhanced Russell Graph measure was introduced for determining closest targets, based on Mixed Integer Linear Programming. Regarding papers that have applied genetic algorithms to solve this type of models, the approach defined in Aparicio et al. (2007) has been recently studied from a metaheuristic perspective (Benavente et al. 2014; López-Espín et al. 2014; González et al. 2015). In Benavente et al. (2014) and López-Espín et al. (2014) heuristics were used to generate valid solutions for a subset of restrictions of Aparicio et al.'s problem, while in González et al. (2015) all constraints are incorporated, heuristics are improved, and new ones are developed, so initial populations of solutions satisfying all the constraints are generated. More recently, González et al. (2016) have taken up where González et al. (2015) left off in the application of metaheuristics and have improved previous heuristics for the generation of valid solutions, seeking also a lower execution time. In addition, a parameterized scheme was introduced working with the initial population of valid and non-valid solutions to generate more valid solutions and to improve all of these solutions to obtain the best fit possible. All these findings were also illustrated by numerical experiments.

This work tries to find the best possible solution to the proposed problem. To this end, two heuristic methods have been developed, in an attempt to find the largest possible number of valid solutions. A solution is considered valid if it meets all the constraints of the problem. Also, this solution is better when the best fit is greater. After obtaining a set of solutions, different metaheuristics are used to improve the first generation of solutions. These metaheuristics include

functions of improvement solutions, combination functions and mutation functions. The improvement functions try to transform the invalid solutions into valid ones, and valid solutions into better valid solutions. The crossover function combines several solutions trying to create a new one, inheriting all the valid qualities from the above solutions. Finally, a mutation function is used to impede stagnation in a local optimum, and have more space to explore solutions. These metaheuristics have certain parameters that specify the operation of each of the previously described functions. These metaheuristics are included in a hyper-heuristic, which attempts to find the best metaheuristic, by training various configurations of these various problems, to find one that achieves the best fit in all types of problems.

After the introduction, Sect. 9.2 will describe the mathematical problem that has been proposed. Having established the problem with all its constraints, section three will address how to find solutions to this problem using heuristic methods. In point three a parameterized scheme of metaheuristics is used to improve all the solutions previously created. Finally, several experiments have been performed to illustrate all the theoretical concepts around the work.

The remainder of the chapter is organized as follows: In Sect. 9.2, we will introduce the necessary notation and background. In particular, we will specify the DEA model that we want to solve. In this way, Sect. 9.3 will address how to find solutions for this problem using heuristic methods. In Sect. 9.4, a parameterized scheme of metaheuristics will be used to improve all the solutions previously defined. Section 9.5 is devoted to performing a comparison between a hyper-heuristic and the results obtained when some pure metaheuristics are applied. Several experiments will be carried out to illustrate all the theoretical concepts of the work. In Sect. 9.7, we present the conclusions.

In order to clarify all concepts provided during this introduction, Fig. 9.1 shows an explanatory diagram where all phases of the work and the techniques used is offered. As can be seen, a hyper-heuristic is found in the high level, where the numerical values of the parameters are determined. These parameters will be later used within a metaheuristic scheme, which consists of: Initialization, improving elements, crossover and mutation. Depending on the values of these parameters, the metaheuristic scheme will approximate more or less to the standard schemes, as genetic algorithms, scatter search or any other. Finally, within the initialization step, there are two heuristic methods through which it is attempted to generate the largest number of solutions possible, where, only those that fulfill all constraints, will be considered as valid. All internal stages of each level will be explained in the following sections of this work.

We also emphasize that both the hyper-heuristic and metaheuristic are independent of the problem, while the heuristic in the initialization is specifically designed for this particular problem. The idea of using a hyper-heuristic as the top level is an advantage for testing the maximum number of possible metaheuristic schemes, in an attempt to find a valid general method for all possible problems.

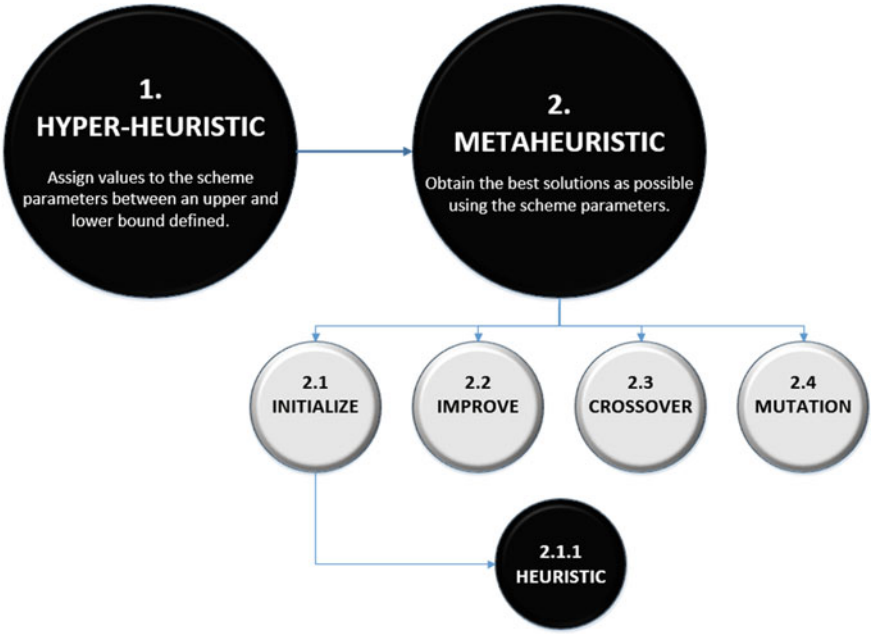


Fig. 9.1 Execution order

9.2 Data Envelopment Analysis and the Problem to Be Solved

Data Envelopment Analysis involves the use of Mathematical Programming to construct a non-parametric piecewise surface over the data in the input-output space. Technical efficiency measures associated with the performance of each DMU are then calculated relative to this surface, as a distance from it.

Now, we introduce some notation. Let us assume that data on m inputs and s outputs for n DMUs are observed. For the j-th DMU these are represented by $x_{ij} \geq 0, \quad i = 1, \ldots, m$ and $y_{rj} \geq 0, \quad r = 1, \ldots, s$.

Basic DEA models are CCR (Charnes et al. 1978) and BCC (Banker et al. 1984). Both models are based on radial projections to the production frontier. However, many other approaches give freedom to the projection so the final efficient targets do not conserve the mix of inputs and outputs. One of these approaches is the traditional Enhanced Russell Graph measure (Pastor et al. 1999), which can be calculated for DMU $k, \quad k = 1, \ldots, n$, as follows:

$$\text{Min} \quad \frac{\frac{1}{m}\sum_{i=1}^{m}\theta_{ik}}{\frac{1}{s}\sum_{r=1}^{s}\phi_{rk}}$$

s.a:

$$\begin{aligned}
&\sum_{j=1}^{n}\lambda_{jk}x_{ij} = \theta_{ik}x_{ik} && i = 1,\ldots,m \\
&\sum_{j=1}^{n}\lambda_{jk}y_{rj} = \phi_{rk}y_{rk} && r = 1,\ldots,s \\
&\theta_{ik} \le 1 && i = 1,\ldots,m \\
&\phi_{rk} \ge 1 && r = 1,\ldots,s \\
&\lambda_{jk} \ge 0 && j = 1,\ldots,n
\end{aligned} \tag{9.1}$$

Equation (9.1) can be easily transformed into the following linear programming model (see Pastor et al. 1999).

$$\text{Min} \quad \beta_k - \frac{1}{m}\sum_{i=1}^{m}\frac{t_{ik}^{-}}{x_{ik}}$$

s.a:

$$\begin{aligned}
&\beta_k + \tfrac{1}{s}\sum_{r=1}^{s}\frac{t_{rk}^{+}}{y_{rk}} = 1 \\
&-\beta_k x_{ik} + \sum_{j=1}^{n}\alpha_{jk}x_{ij} + t_{ik}^{-} = 0 && i = 1,\ldots,m \\
&-\beta_k y_{rk} + \sum_{j=1}^{n}\alpha_{jk}y_{rj} - t_{rk}^{+} = 0 && r = 1,\ldots,s \\
&\beta_k \ge 0 \\
&t_{ik}^{-} \ge 0 \\
&t_{rk}^{+} \ge 0 && r = 1,\ldots,s \\
&\alpha_{jk} \ge 0 && j = 1,\ldots,n
\end{aligned} \tag{9.2}$$

The Enhanced Russell Graph measure, defined as the optimal value of the Eq. (9.1), satisfies several interesting properties. However, it presents a drawback that is also common to other traditional measures in DEA. In particular, the traditional Enhanced Russell Graph measure yields efficient targets that are far from DMU k. The objective function in Eq. (9.1) is normally minimized. Therefore, in order to determine the closest efficient targets, it would seem sufficient to change

"min" for "max" in Eq. (9.1). However, this is not so. In this case, we could show that the targets generated by the model would not be technically efficient, but inefficient (see Färe et al. 1985) and, therefore, could not serve as a valid benchmark for the assessed DMU. In Farë et al. (1985), this problem is solved by resorting to Mixed Integer Linear Programming. In the case of DMU k, the model to be solved would be:

$$\text{Max} \quad \beta_k - \frac{1}{m}\sum_{i=1}^{m}\frac{t_{ik}^-}{x_{ik}}$$

s.a:

$$\begin{aligned}
&\beta_k + \tfrac{1}{s}\sum_{r=1}^{s}\frac{t_{rk}^+}{y_{rk}} = 1, && && (c.1)\\
&\sum_{j=1}^{n}\alpha_{jk}x_{ij} = \beta_k x_{ik} - t_{ik}^-, && i = 1,\ldots,m && (c.2)\\
&\sum_{j=1}^{n}\alpha_{jk}y_{rj} = \beta_k y_{rk} + t_{rk}^+, && r = 1,\ldots,s && (c.3)\\
&-\sum_{i=1}^{m}v_{ik}x_{ij} + \sum_{r=1}^{s}\mu_{rk}y_{rj} + d_{jk} = 0, && j \in E && (c.4)\\
&v_{ik} \geq 1, && i = 1,\ldots,m && (c.5)\\
&\mu_{rk} \geq 1, && r = 1,\ldots,s && (c.6)\\
&d_{jk} \leq M\, b_{jk}, && j \in E && (c.7)\\
&\alpha_{jk} \leq M(1 - b_{jk}), && j \in E && (c.8)\\
&b_{jk} \in \{0,1\}, && j \in E && (c.9)\\
&\beta_k \geq 0, && && (c.10)\\
&t_{ik}^- \geq 0, && i = 1,\ldots,m && (c.11)\\
&t_{rk}^+ \geq 0, && r = 1,\ldots,s && (c.12)\\
&d_{jk} \geq 0, && j \in E && (c.13)\\
&\alpha_{jk} \geq 0, && j \in E && (c.14)
\end{aligned} \tag{9.3}$$

In Eq. (9.3), the uses of a "big M" in (c.7) and (c.8) includes a weakness. These constraints allow us to link d_{jk} to α_{jk} by means of the binary variable b_{jk}. The value of M can be calculated if and only if all the facets that define the DEA technology are previously determined. Unfortunately, the identification of all these facets is a combinatorial NP-hard problem (Frei and Harker 1999; Cherchye and Van Puyenbroeck 2001; Gonzalez and Alvarez 2001). In this paper, we apply several heuristics to solve Eq. (9.3) under the command of a parameterized scheme of metaheuristics. In order to check the viability of our approach, the results are

compared with those obtained from the determination of all the facets of the DEA frontier using a set of simulated numerical examples.

9.3 Heuristic Models

A heuristic technique is any approach to problem solving, learning, or discovery that employs a practical methodology but it not guaranteed to be optimal or perfect, but sufficient for the immediate goals. When an optimal solution is not known, heuristic methods can be used to speed up the process of finding a satisfactory solution. Heuristics can be shortcuts to obtain a satisfactory solution when the optimal solution is unknown.

The main objective of this work is to get results as close as possible to the optimum in problems where we only know the size of the problem and the values of input and output. For this, heuristic methods are used to find, in the first instance, valid solutions that meet the requirements of the problem. A solution will only be valid if it satisfies all the constraints included in the problem.

The solution of each equation is composed of $\beta, \alpha, t^+, t^-, \mu, v$ and d. In order to meet the 14 constraints of Eq. (9.3), a high number of tests and combinations have been developed, searching for the best combination of values and operators that generate a greater number of initial valid solutions. Two methods to generate the initial population of solutions are combined. First, method 1 is used, and if non valid solutions are obtained, method 2 is used in order to try and find a valid solution. The second method has a higher computational cost than the first one, and so it is only used when the first method fails.

Regarding additional notation, r is the subscript for outputs, i for inputs and j for DMUs. The assessed DMU is specified with k.

The scheme of the two heuristic methods used is explained below:

Method 1

1. The process starts generating $b_{jk} \forall j$ based on c.9, where k is the DMU analyzed, and it is comprised between 1 and n. The number of $b'_{jk}s$ equals 0 is greater than or equal to s and lower than or equal to m. The positions of the 0 are randomly generated. So, the values of α_{jk} and $d_{jk} \forall j$ can be calculated by mean of systems of equations in the next steps (α_{jk}, d_{jk} and b_{jk} are related through c.7 and c.8).
2. $t_{rk}^+ \forall r$ and β_k are generated using Algorithm 1 in order to satisfy c.1. t_{rk}^+ are generated randomly between 0 and 1. Next, β_k is obtained using c.1. If β_k is lower than 0, r is randomly generated between 0 and m and t_{rk}^+ is decreased dividing by a parameter p. This parameter p is modified in the experiments to find the value that generates better solutions ($p = 1.05$). If β_k is greater than 1, t_{rk}^+ is increased multiplying by the same parameter. The process continues until $0 < \beta_k < 1$. All of this is shown schematically in Algorithm 1.

Algorithm 1: Generation of t_{rk}^{+} and β_k.

Require: $Y \in R^{+^{s\,x\,n}}, DMU\ k, p$
Ensure: $\forall r, t_{rk}^{+} \in R^{+}, 0 < \beta_k < 1$
  Generate $\forall r, t_{rk}^{+}$ randomly between 0 and 1
  Obtain β_k using c.1.
  While $\beta_k \leq 0$ ***OR*** $\beta_k \geq 1$ **do**
    If $\beta_k \leq 0$ **then**
      Generate *r* randomly, and $t_{rk}^{+} = t_{rk}^{+}/p$
    Else
      Generate *r* randomly, and $t_{rk}^{+} = t_{rk}^{+} * p$
    End if
    Obtain β_k using c.1.
  End while

3. For each $b_{jk} \neq 0$, α_{jk} must be zero. The other values for vector α_{jk} are obtained by solving the system of equations from c.3 (Algorithm 2).

Algorithm 2: Calculate α_{jk} to satisfy c.3.

Require: $Y \in R^{+^{s\,x\,n}}, t_{rk}^{+} \in R^{+}, b_{jk}, \beta_k, DMU\ k$
Ensure: $\forall j, \alpha_{jk} \in R^{+}$
  If $\left|\{b_{jk} = 0 / j = 1, \ldots, n\}\right| - s > 0$
    For j= 1, … , ***n*** **do**
      If $b_{jk} = 0$, *then* α_{jk} is generated randomly between [0,1].
    End For
  End If
  Calculate all the others α_{jk} using c.3.

4. $t_{ik}^{-} \forall i$ are calculated using c.2 by solving the system of equations.
5. Finally, $v_{ik} \forall i$, $\mu_{rk} \forall r$, $d_{jk} \forall j$ are calculated using c.4. The number of d_{jk} equal to 0 must match with the number of α_{jk} different from 0. Because of that, the equations where $d_{jk} = 0$ are used to calculate the other characteristics (v_{ik} and μ_{rk}). After that, the others d_{jk} are calculated. Algorithm 3 is used to obtain all the variables described here.

Algorithm 3: Calculate μ_{rk}, v_{ik} and d_{jk} to satisfy c.4.

Require: $Y \in R^{+^{s\,x\,n}}$, $X \in R^{+^{m\,x\,n}}$, $DMU\ k$
Ensure: $\forall j, d_{jk} \in R^{+}$; $\forall r, \mu_{rk} \in R^{+}$; $\forall i, v_{ik} \in R^{+}$
If $\Delta = \left|\{\alpha_{jk} \neq 0 / j = 1, \ldots, n\}\right| - s = 0$ **then**
Generate v_{ik} randomly $\forall i$
Calculate μ_{rk}, $\forall r$ using equations of indices j where $\alpha_{jk} \neq 0$ in c.4
End if
If $\Delta = \left|\{\alpha_{jk} \neq 0 / j = 1, \ldots, n\}\right| - s > 0$ **then**
$\sigma = m + s - \left|\{\alpha_{jk} \neq 0 / j = 1, \ldots, n\}\right|$
For $i = 1 \ldots . \sigma$ **do**
Generate $\forall i, v_{ik}$ randomly
End for
Calculate $\forall r, \mu_{rk}$ and $v_{ik}, i = \sigma + 1, \ldots, m$ using equations of indices j where $\alpha_{jk} \neq 0$ in c.4
End if
With μ_{rk} and v_{ik} calculate $\forall j, d_{jk}$ using c.4.

Method 2

This method is used when the first method fails to find a valid solution. The heuristic generates new random solutions and tries to make them valid. This method has a higher computational cost, but is more efficient that method 1.

1. $b_{jk} \forall j$ are randomly generated based on c.9. In this method, the number of $b'_{jk}s$ equals 0 is greater than or equal to 1 and lower than or equal to m.
2. For all j = 1, …, n such that $b_{jk} = 0, \alpha_{jk}$ is generated randomly in (0,1).
3. For all j = 1, …, n such that $b_{jk} = 0, \alpha_{jk}$ is modified using Algorithms 4 and 5 in order to satisfy c.1, c.2, c.3, c.11 and c.12. In Algorithm 4 the maximum value of $\beta_k x_{ik}$ in c.2 is equal to $x_{ik} \forall i$, since β_k is between 0 and 1. Then, $\sum_{j=1}^{n} \alpha_{jk} x_{ij}$ must be lower than $x_{ik} \forall i$ in order to satisfy c.11. Therefore, the α with least effect in c.3, denoted by $\alpha_{j_0 k}$ must be decreased. To decrease this value, a constant q_0 has been used, multiplying this constant by $\alpha_{j_0 k}$. In practice, the value of 0.95 has been used to decrease the original value by 5%. Otherwise, $\sum_{j=1}^{n} \alpha_{jk} y_{rj}$ must be greater than y_{rk} $\forall r$ in order to satisfy c.12. In the same way, the α with least effect in c.2 is increased. To increase this value, a constant q_1 has been used, multiplying this constant by $\alpha_{j_0 k}$. This parameter q_1 has been established experimentally at 1.05, increasing the original value by 5%. Algorithm 5 has been developed considering c.1 and c.3. In it, the α with least effect in c.2 and c.3, $\alpha_{j_0 k}$, is calculated in order to satisfy c.1.

Algorithm 4: Adjust α_{jk} to satisfy c.2 and c.3.

Require: $Y \in R^{+^{s\,x\,n}}$, $X \in R^{+^{m\,x\,n}}$, $DMU\ k$
Ensure: $\forall r, t_{rk}^{+} \in R^{+}$; $\forall i, t_{ik}^{-} \in R^{+}$; $\forall j, \alpha_{jk} \in R^{+}, 0 < \beta_k < 1$
For $j = 1 \ldots . m$ **do**
If $x_{ik} < \sum_{j=1}^{n} \alpha_{jk} x_{ij}$ **then**
Find $j_0 \,/\, \frac{1}{m}\sum_{i=1}^{m} x_{ij_0} - \frac{1}{s}\sum_{i=1}^{s} y_{ij_0} = max_{j=1,\ldots,n} \{\frac{1}{m}\sum_{i=1}^{m} x_{ij} - \frac{1}{s}\sum_{i=1}^{s} y_{ij}\}$
$\alpha_{j_0 k} = \alpha_{j_0 k} * q_0$
End if
End for
For $j = 1 \ldots . s$ **do**
If $y_{rk} > \sum_{j=1}^{n} \alpha_{jk} y_{rj}$ **then**
Find $j_0 \,/\, \frac{1}{s}\sum_{i=1}^{s} y_{ij_0} - \frac{1}{m}\sum_{i=1}^{m} x_{ij_0} = max_{j=1,\ldots,n} \{\frac{1}{s}\sum_{i=1}^{s} y_{ij} - \frac{1}{m}\sum_{i=1}^{m} x_{ij}\}$
$\alpha_{j_0 k} = \alpha_{j_0 k} * q_1$
End if
End for
$\forall j$ adjust α_{jk} with algorithm 5.
Adjust β_k to satisfy c.11. and c.12. (step 4)
Obtain t_{rk}^{+} $\forall r$ and t_{ik}^{-} $\forall i$ using c.2. and c.3. (step 4)

Algorithm 5: Adjust α_{jk} to satisfy c.1.

Require: $Y \in R^{+^{s\,x\,n}}$, $X \in R^{+^{m\,x\,n}}$, $DMU\ k$, $\alpha_{jk}\ \forall j$
Ensure: $\forall j, \alpha_{jk} \in R^{+}$
$\forall j, 1 \le j \le n, p_j = \sum_{r=1}^{s} y_{rj}/y_{rk}$
repeat
Find $j_0 \,/\, \frac{1}{s}\sum_{i=1}^{s} y_{ij_0} + \frac{1}{m}\sum_{i=1}^{m} x_{ij_0} = min_{j=1,\ldots,n} \{\frac{1}{s}\sum_{i=1}^{s} y_{ij} - \frac{1}{m}\sum_{i=1}^{m} x_{ij}\}$
$sum = \sum_{j=1,\ldots,n, j \neq j_0} \alpha_{jk} p_j$
$\alpha_{j_0 k} = \frac{s - sum}{p_{j_0}}$
If $\alpha_{j_0 k} \le 0$ **then**
$\alpha_{j_0 k} = \alpha_{j_0 k} * q_0$
End if
Until $\alpha_{j_0 k} > 0$

---

4. Iteratively adjust β_k to satisfy c.2, c.3, c.11 and c.12 in the following way. If c.11 is violated, then β_k is increased by a factor to satisfy c.2 and c.11. Otherwise, if c.12 is violated, then β_k is decreased by a factor to satisfy c.3 and c.12. This factor is decreased in each iteration for a finer adjustment. In each iteration $t_{rk}^{+} \forall r$ and $t_{ik}^{-} \forall i$ are obtained using c.2 and c.3.

5. Finally, $v_{ik} \forall i, \mu_{rk} \forall r, d_{jk} \forall j$ are calculated using c.4. The number of d_{jk} equal to 0 is the same as the number of α_{jk} different from 0. Algorithm 3 is used to obtain all the variables.

9.4 Parameterized Scheme of Metaheuristics

In a previous work, González et al. (2015), the solution to the problem was addressed with a genetic algorithm. Here, however, the genetic algorithm has been replaced by a parameterized metaheuristic scheme. This offers the possibility of analyzing a large number of different metaheuristics with the aim of finding one that is satisfactory for the problem in question. The parameterized scheme comprises a skeleton (Algorithm 6) with six basic functions: Initialize, EndCondition, Selection, Combine, Improve and Include.

To analyze how the parameterized scheme can be applied to our problem, three basic metaheuristics are considered (Almeida et al. 2013a): Greedy Randomized Adaptive Search Procedure (GRASP), Scatter Search (SS) and Genetic Algorithms (GA). They are implemented within the scheme, and the inclusion of parameters allows us to experiment with the three basic metaheuristics and with different types of hybridizations. As shown in Algorithm 6, a group of initial solutions are generated, and while the end condition is not satisfied, they are selected, combined and improved.

Algorithm 6: Parameterized metaheuristic scheme.

```
Require: S, ParamINI, ParamEND, ParamSEL, ParamCOM, ParamIMP, ParamINC.
    Initialize (S, ParamINI)
    while (not EndCondition(S, ParamEND)) do
        SS = Select (S, ParamSEL)
        SS1 = Combine (SS, ParamCOM)
        SS2 = Improve (SS1, ParamIMP)
        S = Include (SS2, ParamINC)
    end while
```

The scheme functions in Algorithm 6 are explained below.

9.4.1 Initializing

A series of initial elements are generated through a set of heuristics, based on the two methods previously discussed. At the end of this process, all valid and invalid elements are improved with certain intensification. When all the elements have been improved, a certain pre-specified number of them will be included in the final group. The parameters used in this first step are: INEIni, FNEIni, PEIIni, IIEIni.

INEIni: Initial Number of Elements. This parameter specifies the number of elements that will be created in the first-generation.
FNEIni: Final Number of Elements. This parameter specifies the number of elements that are used to obtain better solutions with the subsequent functions.
PEIIni: Percentage of Elements to Improve. A percentage of elements are improved. If the solution is a valid solution, the improvement tries to enhance it. On the other hand, if the solution is an invalid solution, the improvement tries to make it valid. This parameter specifies the percentage of elements that are going to be improved after.
IIEIni: Intensification of the Improvement of Elements. This last parameter specifies the number of times that an improvement is applied to a solution.

At the end of this step, a set of initial elements are generated and improved.

9.4.2 End Condition

The parameters used in this second step are: MNIEnd, NIREnd.

MNIEnd: Maximum Number of Iterations. This parameter specifies how many times the "while" loop from the Algorithm 4 is repeated for each DMU.
NIWIEnd: Number of Iterations Without Improving the best solution. After some iterations while keeping the best element fixed, the crossover and improvement operations are completed and passed to the following DMU.

Other end conditions are contemplated and are implemented in the program code. The other criteria make reference to the objective function value. If the fitness function value is higher than 1 during more than INEIni/10 consecutive times, or if the fitness function value is higher than 1 more than INEIni/2 twice non-consecutively the program will end.

9.4.3 Selecting

Generated elements are classified into two different groups; valid and invalid solutions. When all the elements are classified into a group, a number of them are selected for further combination. With these selected solutions, the crossover function generates new elements than inherit the characteristics of their predecessors. The parameters used in this third step are: NBESel, NWESel.

NBESel: Number of Best Elements Selected. This parameter specifies how many valid elements are going to be used in the crossover function. The elements of this group have previously been ordered by the value of the objective function from highest to lowest. As a rule, the best elements of this group are selected.
NWESel: Number of Worst Elements Selected. This parameter performs the same operation as above, with the difference that the elements selected come from the invalid group.

Once we have all the desired items stored in their respective groups, we proceed to send it to the crossover and mutation functions.

9.4.4 Combining

New elements will be created using the two groups in the selection step. To make combinations, two elements are chosen randomly from the selected group. Their characteristics are combined using two different methods: Arithmetic and By Points.

The crossover of elements is based on the natural evolution itself. From two elements selected from a large group of specimens, we combine their characteristics to generate a new specimen which shares the characteristics of its predecessors, and generates new ones from them.

Because each individual has components of five types (β, t^+, t^-, μ and v), each combination will work with only one of these types. Each crossover uses its own characteristic to be combined and create new ones.

Arithmetic Crossover A characteristic is randomly selected from the group of five specified before. The values from this characteristic are combined with a mathematical operation (addition, subtraction or average).

Different operations have been tested depending on the particular characteristic considered. The new values are included in the new solution, and the other characteristics of the solution are recalculated, in order to make it valid.

For example, if a chromosome has the characteristic $t_1^+ [0, 2, 1, 4]$ and another $t_2^+ [1, 1, 1, 2]$, the result will be $t_3^+ [0 + 1, 2 + 1, 1 + 1, 4 + 2] = [1, 3, 2, 6]$. That is an example with an operation of addition. Another example can be an average operation:

$t^+$from Chromosome 1

0	2	1	4

t^+ from Chromosome 2

1	1	1	2

Average [Chromosome1, Chromosome2]

(0+1)/2	(2+1)/2	(1+1)/2	(4+2)/2

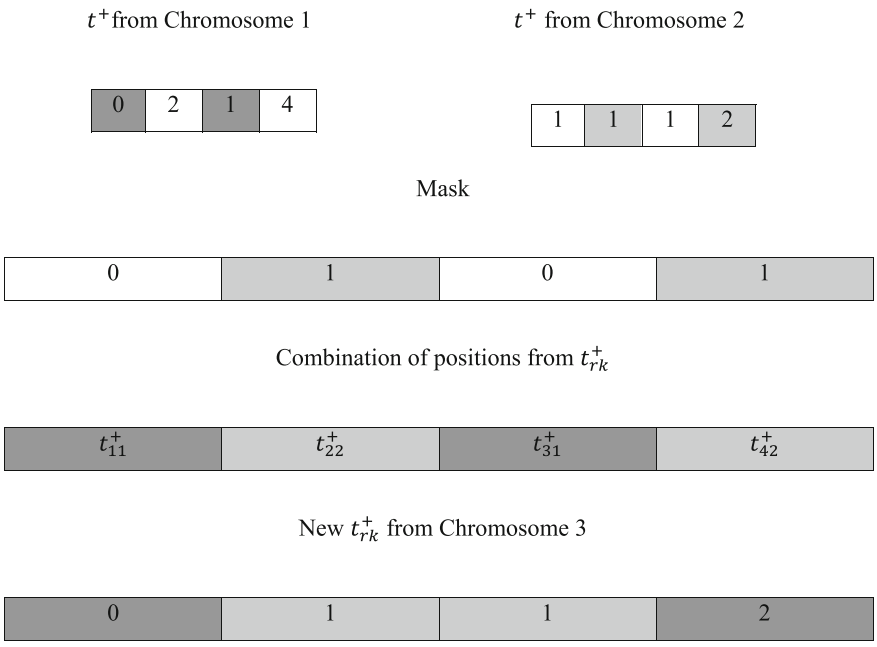
$t^+$from Chromosome 3 (Average[Chrom1, Chrom2]

0.5	1.5	1	3

Point Crossover A characteristic is selected randomly like before. All the values of the new solution are taken from one of the selected solutions, and one of them is replaced by the corresponding value of the other solution. A binary mask algorithm could be used to change more than one value in the new solution, but satisfactory solutions are obtained with this simple approach. The other characteristics of the solution are recalculated.

This method is really interesting when the characteristic has a large number of values, because in this case, the mask will have more changes inside. When it is a short one, the mask only uses a few cutoff points.

Here is an example with the same characteristic as before. A randomly generated mask is generated to combine the features of both chromosomes.



The parameters used in this fourth step are: PBBCom, PWWCom and PBWCom.

PBBCom: Number of combinations between elements with better fitness. This parameter indicates the number of combinations between valid elements.
PWWCom: Number of combinations between elements with worse fitness. This parameter indicates the number of combinations between invalid elements.
PBWCom: Number of combinations between valid and invalid elements. This parameter indicates the number of combinations between valid and invalid elements.

As mentioned above, five variables are used to make crosses between chromosomes (β, t^+, t^-, μ and v). More than one characteristic can be used for a single crossover, and different crossovers can use the same characteristic but using different operations. All the possible crossovers implemented are shown below:

1. β: The average of β of the selected items is obtained, and the value is modified by adding a small randomly generated value.
2. β: The average of β is obtained, and is modified by subtracting a value.
3. μ_{rk}: An arithmetic crossover is used on μ_{rk}.
4. t^+_{rk}: An arithmetic crossover is used on t^+_{rk}.
5. v_{ik}: A crossover by points is used on v_{ik}.
6. t^+_{rk}: A crossover by points is used on t^+_{rk}.
7. t^-_{ik}: A crossover by points is used on t^-_{ik}.

8. t_{rk}^{+} and v_{ik}: This method is a combination of types 5 and 6.
9. t_{rk}^{+}: An arithmetic crossover with subtraction is used on t_{ik}^{-}, and the initial values of t_{rk}^{+} are substituted by those of t_{ik}^{-}. It has been observed experimentally that this crossover process increases the number of valid solutions.

To determine which crossover is better in the creation of optimal solutions, a high number of iterations have been executed, using a single problem with sizes $m = 3$, $n = 30$, $s = 2$ as an example. The problem is executed 10 times with 100 iterations. In each execution, an optimal solution is found for every DMU and each solution has been created using the different kinds of crossovers. Figure 9.2 shows the percentage of times that each crossover is applied to create an optimal solution.

The variables which are most often involved in the creation of optimal solutions are t_{rk}^{+} and β. The method that achieves the best results uses t_{rk}^{+} (crossover 4) with an arithmetic crossover. Consequently, the metaheuristic is improved with a crossover function in which more probability is assigned to the crossovers with a high percentage as in Fig. 9.2.

Each solution to a particular problem is improved with a crossover function. Using the same problem as the previous experiment, the final optimal solutions after being crossed have been compared to the solutions obtained in the first generation. The objective is to know if the crossover algorithm works for each solution.

A single problem ($m = 3$, $n = 30$, $s = 2$) has been implemented as an example, with the parameters shown in Table 9.1 (the rest are 0):

As Table 9.2 shows, all the elements obtained in the first generation are improved to a greater or lesser extent, by passing through the crossover function. Also, the crossover function obtains solutions very close to the solutions obtained by optimization software. When the crossover function ends, all the solution in the reference set are valid solutions. The optimization software uses the additive

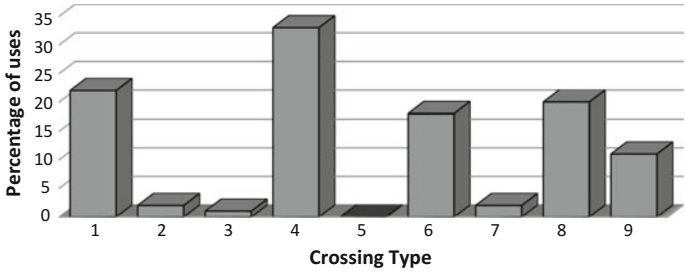


Fig. 9.2 Comparison of the different types of crossover

Table 9.1 Metaheuristic parameters

INEIni	FNEIni	MNIEnd	NIREnd	NBESel	NWESel	PBBCom	PWWCom
100	100	100	50	25	25	100	100

Table 9.2 Fitness comparison

	First generation		After crossover	
DMU	% valid	Fitness	% valid	Fitness
1	100	0.417	100	0.471
2	100	0.280	100	0.290
3	100	0.476	100	0.481
4	100	0.825	100	0.871
6	3	0.908	100	0.908
7	100	0.644	100	0.675
9	100	0.535	100	0.540
11	100	0.484	100	0.485
12	6	0.588	100	0.620
13	100	0.758	100	0.762
14	100	0.171	100	0.211
16	100	0.618	100	0.655
17	100	0.771	100	0.771
18	11	0.952	100	0.952
20	100	0.182	100	0.230
21	100	0.397	100	0.401
22	100	0.170	100	0.171
24	100	0.430	100	0.512
25	100	0.651	100	0.664
26	100	0.056	100	0.058
27	98	0.015	100	0.015
28	48	0.835	100	0.873
29	16	0.920	100	0.920

algorithm to obtain the efficient solutions. Afterwards, the algorithm searches the combinations between them that generate an efficient frontier. Finally, the distances between the infeasible solutions and the frontiers are calculated. This is a NP-hard problem, and this problem needs considerable time to be solved.

9.4.5 Improving the Solution

The improvement is used for all the elements generated in the first generation, and for all the elements generated by the crossover function. There are two types of improvements: one for the valid elements, which try to achieve a better solution by improving fitness, and another one for the invalid elements to try and make them valid. The parameters used in this fifth step are: PEIImp, IEIImp, PEDImp, IIDImp.

PEIImp: Percentage of Elements to Improve by local search. After completing all the established crossover, and the solutions have then become diversified through

mutations, we proceed to try and improve all the valid solutions through a local search. The improvement function is the same as previously used for the elements of the first generation, with the exception that in this case only the valid solutions are improved.

IEIImp: Intensification on the Elements of the Improvement. This parameter specifies the number of times that an improvement is applied to an element. This value is shared by all the improvements in the first generation and the improvements by the crossover function.

PEDImp: Percentage of Elements for Diversification. In order to have various elements in all the solution spaces, a mutation function has been implemented. This function modifies all the internal characteristics from one solution. This is necessary to avoid getting stuck at a local optimum. Furthermore, diversification allows for different parts of the solution space to be explored.

IIDImp: Intensification of the improvement of elements by diversification. This is the same as the other improvements. The new elements that come from a mutation are improved with the same function as before.

9.4.6 Including

In order to have more variety of opportunities to find an optimal solution, all final solutions of each iteration (in a number equal to FNEIni) are used to create new chromosomes in the following iterations. In conclusion, all the FNEIni solutions created at the end of iteration are used as a reference set in the next step.

9.5 Metaheuristic Models

With all these steps, a parameterized scheme of metaheuristic is performed. As we have explained previously, several basic metaheuristics have been used to make a comparison between the different fitness obtained with each of them. To get the best parameter configuration, we perform a hyper-heuristic comparing the results obtained with it, with those obtained with some pure metaheuristics applied directly to the optimization problem. The next table shows the values selected for the parameters for the three basic metaheuristics (Table 9.3).

The most important characteristics of all the basic metaheuristics than we proposed previously are explained below:

Genetic Algorithm: This is a search heuristic that mimics the process of natural selection. Genetic algorithms belong to the larger class of evolutionary algorithms, which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. As we can see in the parameters table, this metaheuristic does not have improvements, and

Table 9.3 Values of the parameters for the three basic metaheuristics considered

Schemes	IINEIni	FNEIni	PEIIni	IEIIni	NBESel	NWESel	PBBCom
GA	100	100	0	0	100	0	50
GR	200	1	100	50	0	0	0
SS	100	20	100	50	10	10	90

Schemes	PWWCom	PEIImp	IEIImp	PEDImp	IDEImp	MNIEnd	NIREnd
GA	0	0	0	10	5	100	25
GR	0	0	0	0	0	100	25
SS	90	100	5	0	0	100	25

only relies on generating a number of initial elements and makes numerous crossovers between them, adding a small number of mutations to diversify the results. In addition, it is important know that this algorithm only uses the valid solutions to make the crossover.

Greedy Randomized Adaptive Search Procedure: This is a metaheuristic algorithm commonly applied to combinatorial optimization problems. GRASP typically consists of improvements in solutions through a local search. In the table we can observe how this metaheuristic generates a lot of elements at the beginning, and only saves the best of them. Then, the algorithm tries to improve this unique solution with a hard improvement.

Scatter Search: This algorithm is based on generating initial elements, and only keeps a few elements in a reference set. Subsequently, it attempts to improve the valid and invalid solutions. Then it selects a set of these solutions to generate combinations together. The Scatter Search algorithm is a derivative of the genetic algorithm, where all elements are used to obtain better solutions with the crossover method. It also includes a crossover function improvement, to try to improve these last elements. Therefore, it is a genetic algorithm without diversification, which uses all sorts of elements to make combinations, and improving all the elements that are generated.

Apart from these basic metaheuristics, a huge number of combinations/hybridizations can be considered simply by selecting different values for the parameters. The best metaheuristic from those obtained with the parameterized scheme could be obtained by generating all the possible combinations of the parameters and by applying them to some small training problems. In this way, the metaheuristic (given by the values of the parameters) which gives the best results for the training set can be considered a satisfactory metaheuristic for the problem in question. The number of possible combinations of the parameters in the parameterized metaheuristic scheme is huge, and the problem for obtaining the best metaheuristic for the training set is an optimization problem. So, it is a suitable problem for metaheuristics. A hyper-heuristic can be developed as a metaheuristic searching for satisfactory metaheuristics, and can be developed over the parameterized metaheuristic scheme. The hyper-heuristic uses the same parameterized scheme. The number of initial metaheuristics and combinations/hybridizations between them are

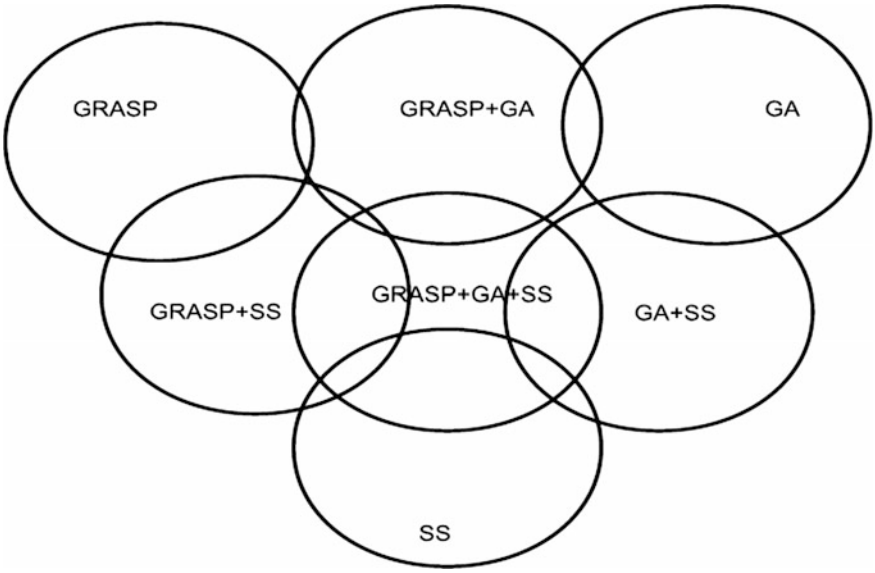


Fig. 9.3 Space of metaheuristics obtained by a combination of the basic metaheuristics

specified and a number of metaheuristics are selected for successive iteration. The combination uses the same algorithm as the previous crossover and, in this case, creates a new one by combining the parameters of the selected metaheuristics. The new metaheuristics are improved by increasing a random parameter. The increase of the parameter is specified with an intensification parameter.

The space of metaheuristics where the hyper-heuristic searches is shown in Fig. 9.3.

9.6 Experimental Results

Experiments were conducted to analyze the effectiveness of the heuristics, metaheuristics and hyper-heuristics developed. The number of valid solutions obtained with the two heuristics in Sect. 9.3 is initially compared. Then, the fitness values obtained with the basic metaheuristics considered are compared with those of hybrid metaheuristics and hyper-heuristics. Finally, the fitness function and the time used by a satisfactory metaheuristic are compared with those from an optimization tool such as CPLEX.

The system used in the experiments is a NUMA node with 4 Intel Nehalem-EX EC E7530 hexa-cores, with 24 cores, at 1.87 GHz and 32 GB of RAM. A parameterized shared-memory metaheuristic scheme (Almeida et al. 2013b) could be used to accelerate the execution time, but the number of experiments is

Table 9.4 Percentage of valid solutions and average of the objective functions for all the DMUs, obtained with the different heuristics when varying the problem size

No	*m*	*n*	*s*	Method 1		Method 2		Method 1 + Method 2	
				% valid	Fitness	% valid	Fitness	% valid	Fitness
1	2	50	1	**100.000**	**0.155**	11.292	0.143	**100.000**	**0.155**
2	3	30	2	81.217	0.498	17.261	0.311	**82.391**	**0.498**
3	4	30	2	83.667	0.436	15.167	0.346	**88.722**	**0.474**
4	4	30	3	74.000	0.523	10.833	0.246	**75.500**	**0.524**
5	5	30	3	24.833	0.134	12.333	0.228	**33.500**	**0.295**
6	6	30	4	4.714	0.135	6.000	0.368	**9.143**	**0.369**

Bold values make reference to the highest values in each row

high and the shared-memory has been exploited with simultaneous executions of the sequential scheme.

The generation of valid solutions for our problem is a difficult task. In previous works the problem associated with obtaining valid solutions for 9 (Benavente et al. 2014) and 13 (González et al. 2015) of the 14 constraints in equation 2 was studied. The two heuristics presented in Sect. 9.3 are used for the generation of valid solutions while fulfilling all the constraints.

Fitness and the percentage of valid solutions generated with the two heuristics are compared in Table 9.4.

When the problem size increases, the number of valid solutions decreases. The first heuristic generates more valid solutions than the second one, and the combination of two heuristics increases the number of valid solutions and the fitness obtained. More valid solutions could be generated with more iteration, but the execution time would increase. A fixed number of iterations are configured using a parameterized scheme of metaheuristics in the next step. Figure 9.4 shows the

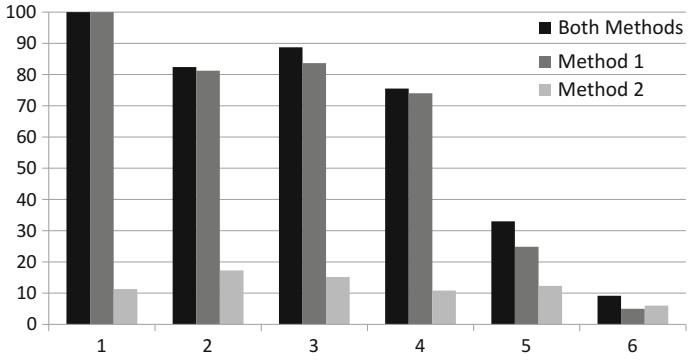


Fig. 9.4 Comparison between the valid solution obtained by using the heuristics

Table 9.5 Metaheuristic values for each problem

m	*s*	*n*	Hyper-heuristic	GA	SS	GR	CPLEX
2	1	50	**0.170**	0.169	**0.170**	**0.170**	0.170
3	2	30	**0.343**	0.334	0.338	0.324	0.383
4	2	28	**0.529**	0.512	**0.529**	0.523	0.548
4	3	20	**0.336**	0.302	0.326	0.316	0.391
5	3	20	**0.456**	0.394	0.380	0.348	0.495

Bold values make reference to the highest values in each row

number of valid solutions generated by each method. Method 1 generates more solutions than method 2, but this last one provides support for method 1. When they work together, the number of valid solutions increases. The horizontal axis makes reference to the problems specified in Table 9.5.

The difference between both methods decreases when the problem size grows. Method 2 is a good support for method 1 for small-sized problems, but for larger-sized problems, the second method works better than the first. For that, both methods are considered and applied to the search of valid solutions. When the problem size increases, the amount of valid solutions obtained in a first generation decreases.

Figure 9.5 compares the fitness obtained with the three basic metaheuristics (GRASP, Genetic Algorithm and Scatter Search) whose parameter values are shown in Table 9.3. Results are shown for small problems of different sizes, and are the average of 10 executions. Small problems are solved by the exact method (CPLEX) obtaining the optimal solution but spending a high computational cost. Finally, the fitness values obtained with the application of a hyper-heuristic to each problem are also shown. The hyper-heuristic generates a hybrid metaheuristic by generating hybrid metaheuristics randomly (a set of parameters for the parameterized scheme) and by combining them with a crossover.

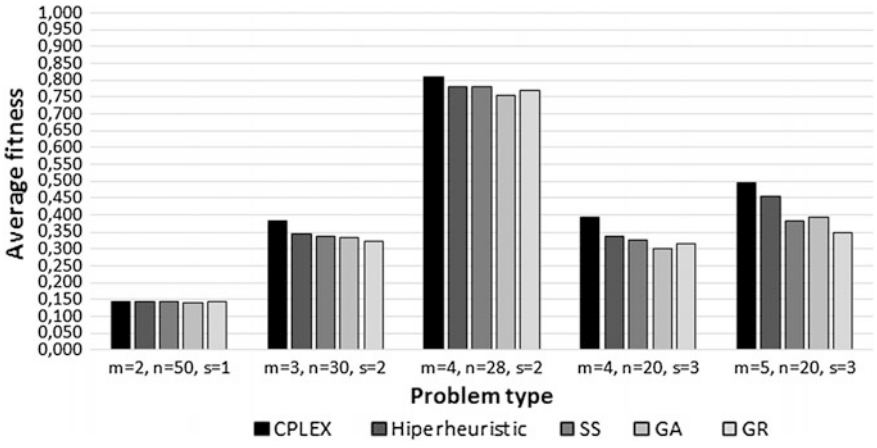


Fig. 9.5 Space of metaheuristics obtained by a combination of the basic metaheuristics

The metaheuristics give fitness values close to the optimum with low execution times. So, they are competitive with exact methods for large problems, where these are impracticable. The application of the hyper-heuristic generates better fitness values than the metaheuristics, especially for larger problems. The hyper-heuristic applied to a problem generates a hybrid metaheuristic (a combination of parameters of the metaheuristic scheme). Table 9.6 shows the values of the parameters of the metaheuristic obtained by the hyper-heuristic for each problem. The parameters obtained with the hyper-heuristic are similar for different problems. A large number of elements is created, and more than 50% of the elements are improved. The number of selected elements does not have a great influence. In contrast, the number of crossovers should be high. The improvement and end parameters vary widely. Some problems improve many elements; others have more mutations.

The hyper-heuristic could be trained with small problems to generate a general metaheuristic with satisfactory behavior for those problems and could work satisfactorily for other problems. The metaheuristics for each problem size can be combined in different ways. For example, each parameter of the metaheuristic can take the highest value from each parameter (Table 9.7). Table 9.8 shows that the fitness of the hyper-heuristic is improved with the general metaheuristic.

The hybrid metaheuristic combines the characteristics of the basic metaheuristics implemented in the parameterized scheme. Figure 9.6 shows the improvement in the solution for a DMU. The figure compares the fitness in each iteration when a problem is approached with different metaheuristics. The Genetic Algorithm starts with the lowest fitness because no improvements are applied in

Table 9.6 Metaheuristic parameters obtained by the hyper-heuristic for different problem sizes

Problem size	IINEIni	FNEIni	PEIIni	IIEIni	NBESel	NWESel	PBBCom
$m = 2, n = 50, s = 1$	185	67	75	11	20	67	84
$m = 3, n = 30, s = 2$	157	54	60	10	26	56	82
$m = 4, n = 28, s = 2$	170	43	58	8	36	42	73
$m = 4, n = 20, s = 3$	172	32	59	3	72	5	50
$m = 5, n = 20, s = 3$	157	27	92	18	27	27	53

Problem size	PWWCom	PEIImp	IIEImp	PEDImp	IIDImp	MNIEnd	NIREnd
$m = 2, n = 50, s = 1$	81	35	10	52	9	51	20
$m = 3, n = 30, s = 2$	75	60	6	38	8	59	36
$m = 4, n = 28, s = 2$	63	13	4	59	8	72	46
$m = 4, n = 20, s = 3$	55	23	7	41	5	91	12
$m = 5, n = 20, s = 3$	30	53	8	13	8	39	43

Table 9.7 Parameters of the general metaheuristic obtained with the hyper-heuristic

IINEIni	FNEIni	PEIIni	IIEIni	NBESel	NWESel	PBBCom
185	67	92	18	72	67	84
PWWCom	PEIImp	IIEImp	PEDImp	IIDImp	MNIEnd	NIREnd
81	60	10	59	9	91	46

Table 9.8 Comparison between the fitness obtained through the direct application to different problems of the hyper-heuristic and with the general metaheuristic obtained by the hyper-heuristic

Schemes	Hyper-heuristic	General metaheuristic
$m = 2$, $n = 50$, $s = 1$	0.142	0.142
$m = 3$, $n = 30$, $s = 2$	0.343	0.348
$m = 4$, $n = 28$, $s = 2$	0.780	0.782
$m = 4$, $n = 20$, $s = 3$	0.336	0.351
$m = 5$, $n = 20$, $s = 3$	0.456	0.460

the initialization, but the fitness is improved with the crossover. On the other hand, a Scatter Search algorithm starts with a better fitness but has fewer improvements in the crossover function. The Greedy Randomized Adaptive Search Procedure works in a suitable way in the initialization. The metaheuristic generated by the hyper-heuristic combines the features of all these algorithms and, consequently, achieves higher fitness.

The parameterized scheme of metaheuristics was used to generate a satisfactory metaheuristic by training a hyper-heuristic with small problems, and the crossover of the metaheuristic was optimized. Better fitness values were obtained, but the optimum solution was not always reached. Even so, metaheuristics are an alternative to exact methods for large problems, for which the execution time of exhaustive methods is unfeasible.

The execution time of an exact method is compared with that of the improved metaheuristic in Fig. 9.7. For the biggest problems, the metaheuristic is approximately 500 times faster than the exact method, and the difference increases with the problem size.

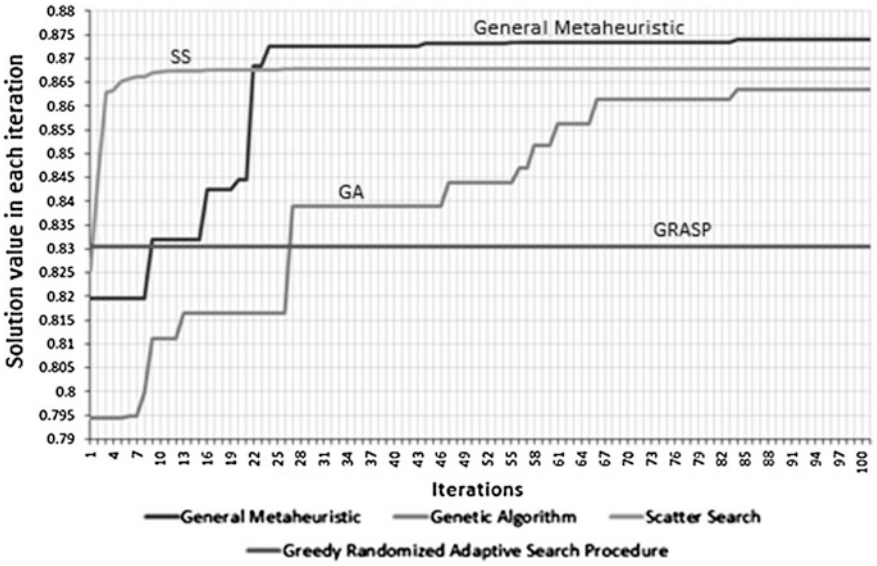


Fig. 9.6 Value of the optimal solution in each iteration

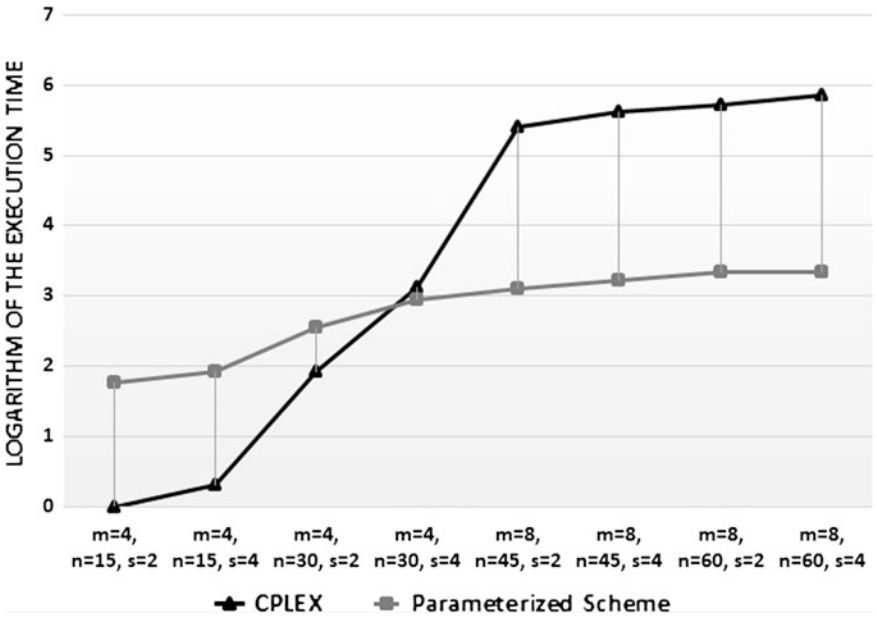


Fig. 9.7 Comparison of the execution time (in seconds and logarithmic scale) between an exhaustive method and the hybrid metaheuristic developed

9.7 Conclusions

The determination of both closest efficient targets and least distance has been an issue of interest in recent DEA literature. This implies the application of a new paradigm where the slacks of the traditional DEA measures are minimized instead of maximized in order to achieve the strongly efficient frontier of the polyhedral production possibility set. Radial measures and Directional Distance Function-type measures are out of the focus of this new approach, since in both cases the projection on the (weakly) efficient frontier is determined following a pre-fixed direction. Nevertheless, the least distance could be applied to the second phase of radial measures in order to get the Pareto-Koopmans frontier.

The implementation of the new approach is clearly more difficult from a computational perspective than that associated with traditional DEA measures (weighted additive models, Enhanced Russell Graph measure, …). This fact has motivated the publication of different approaches trying to implement the problem in a suitable way. In this respect, other heuristics have been previously introduced in the literature. However, the new heuristic proposed here provides more valid solutions satisfying all the constraints in the model and with a lower execution time. A parameterized scheme has been developed working with this initial population of valid and invalid solutions to generate more valid solutions and to improve all of these solutions to obtain the best fitness possible.

Additionally, the hyper-heuristic generated with all of the used basic metaheuristics gives solutions close to the optimum and is competitive with an exact method with a high computational cost, which cannot be used for large problems. A deeper analysis should be made to tune the hyper-heuristic to obtain better solutions with lower execution times.

Overall, the new paradigm applied to the so-called Enhanced Russell Graph measure has been studied. Nevertheless, there are a lot of measures in DEA that can be used in the maximization of technical efficiency or, equivalently, in the minimization of a certain distance. In this way, programming the approach based on metaheuristic algorithms to solve all of them could be seen as an appropriate and interesting future work.

References

Aigner DJ, Chu SF (1968) On estimating the industry production function. Am Econ Rev 58: 826–839

Aigner DJ, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. J Econ 6:21–37

Almeida F, Giménez D, López-Espín JJ, Pérez-Pérez M (2013a) Parameterized schemes of metaheuristics: basic ideas and applications with genetic algorithms, scatter search and GRASP. IEEE Trans Syst Man Cybern 43(3):570–586

Almeida F, Giménez D, López-Espín JJ (2013b) A parameterized shared-memory scheme for parameterized metaheuristics. J Supercomput 58(3):292–301

Amirteimoori A, Kordrostami S (2010) A Euclidean distance-based measure of efficiency in data envelopment analysis. Optimization 59:985–996
Ando K, Kai A, Maeda Y, Sekitani K (2012) Least distance based inefficiency measures on the Pareto-efficient frontier in DEA. J Oper Res Soc Jpn 55(1):73–91
Aparicio J, Pastor JT (2013) A well-defined efficiency measure for dealing with closest targets in DEA. Appl Math Comput 219:9142–9154
Aparicio J, Pastor JT (2014a) On how to properly calculate the Euclidean distance-based measure in DEA. Optimization 63(3):421–432
Aparicio J, Pastor JT (2014b) Closest targets and strong monotonicity on the strongly efficient frontier in DEA. Omega 44:51–57
Aparicio J, Ruiz JL, Sirvent I (2007) Closest targets and minimum distance to the Pareto-efficient frontier in DEA. J Prod Anal 28:209–218
Aparicio J, Mahlberg B, Pastor JT, Sahoo BK (2014a) Decomposing technical inefficiency using the principle of least action. Eur J Oper Res 239:776–785
Aparicio J, Borras F, Ortiz L, Pastor JT (2014b) Benchmarking in healthcare: an approach based on closest targets. In: Emrouznejad A, Cabanda E (ed) Managing service productivity. International series in operations research & management science, vol 215, Springer, Berlin, pp 67–91
Baek C, Lee J (2009) The relevance of DEA benchmarking information and the least-distance measure. Math Comput Model 49:265–275
Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30:1078–1092
Benavente C, López-Espín JJ, Aparicio J, Pastor JT, Giménez D (2014) Closets targets, benchmarking and data envelopment analysis: a heuristic algorithm to obtain valid solutions for the shortest projection problem. In: 11th international conference on applied computing
Briec W (1997) Minimum distance to the complement of a convex set: duality result. J Optim Theory Appl 93(2):301–319
Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444
Charnes A, Cooper WW, Golany B, Seiford L, Stutz J (1985) Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. J Econ 30:91–107
Cherchye L, Van Puyenbroeck T (2001) A comment on multi-stage DEA methodology. Oper Res Lett 28:93–98
Cobb CW, Douglas PH (1928) A theory of production, vol 18, issue 1. In: The American economic review, supplement, papers and proceedings of the fortieth annual meeting of the American economic association, pp 139–165
Cooper WW, Park KS, Pastor JT (1999) RAM: a range adjusted measure of inefficiency for use with additive models, and relations to others models and measures in DEA. J Prod Anal 11:5–42
Färe R, Grosskopf S, Lovell CAK (1985) The measurement of efficiency of production. Kluwer Nijhoff Publishing, Boston
Farrel MJ (1957) The measurement of productive efficiency. J Roy Stat Soc Ser A (General) 120 (3):253–290
Frei FX, Harker PT (1999) Projections onto efficient frontiers: theoretical and computational extensions to DEA. J Prod Anal 11:275–300
Fukuyama H, Maeda Y, Sekitani K, Shi J (2014a) Input-output substitutability and strongly monotonic p-norm least-distance DEA measures. Eur J Oper Res 237:997–1007
Fukuyama H, Masaki H, Sekitani K, Shi J (2014b) Distance optimization approach to ratio-form efficiency measures in data envelopment analysis. J Prod Anal 42:175–186
Fukuyama H, Hougaard JL, Sekitani K, Shi J (2016) Efficiency measurement with a nonconvex free disposal hull technology. J Oper Res Soc (in press)
González E, Alvarez A (2001) From efficiency measurement to efficiency improvement: the choice of a relevant benchmark. Eur J Oper Res 133:512–520

González M, López-Espín JJ, Aparicio J, Giménez D, Pastor JT (2015) Using genetic algorithms for maximizing technical efficiency in data envelopment analysis. Procedia Comput Sci 51 (2015):374–383

González M, López-Espín JJ, Aparicio J, Gimenéz D (2016) Implementing the principle of least action in data envelopment analysis: a parameterized scheme of metaheuristics. Working Paper, University Miguel Hernandez of Elche

Jahanshahloo GR, Hosseinzadeh Lotfi F, Zohrehbandian M (2005) Finding the piecewise linear frontier production function in data envelopment analysis. Appl Math Comput 163:483–488

Jahanshahloo GR, Lotfi FH, Rezai HZ, Balf FR (2007) Finding strong defining hyperplanes of production possibility set. Eur J Oper Res 177:42–54

Jahanshahloo GR, Vakili J, Mirdehghan SM (2012) Using the minimum distance of DMUs from the frontier of the PPS for evaluating group performance of DMUs in DEA. Asia-Pac J Oper Res 29(2):1250010

Koopmans TC (1951) Analysis of production as an efficient combination of activities. In: Koopmans TC (ed) Activity analysis of production and allocation. John Wiley, New York

López-Espín JJ, Aparicio J, Giménez D, Pastor JT (2014) Benchmarking and data envelopment analysis. An approach based on metaheuristics. Procedia Comput Sci 29:390–399

Lozano S, Villa G (2005) Determining a sequence of targets in DEA. J Oper Res Soc 56: 1439–1447

Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb-Douglas production functions with composed error. Int Econ Rev 18:435–444

Pastor JT, Aparicio J (2010) The relevance of DEA benchmarking information and the least-distance measure: comment. Math Comput Model 52:397–399

Pastor JT, Ruiz JL, Sirvent I (1999) An enhanced DEA Russell graph efficiency measure. Eur J Oper Res 115:596–607

Portela MCS, Borges PC, Thanassoulis E (2003) Finding closest targets in non-oriented DEA models: the case of convex and non-convex technologies. J Prod Anal 19:251–269

Tone K (2001) A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res 130:498–509

Part III
Empirical Applications

Chapter 10
Producing Innovations: Determinants of Innovativity and Efficiency

Jaap W.B. Bos, Ryan C.R. van Lamoen and Mark W.J.L. Sanders

Abstract In this chapter, we investigate the knowledge production function, using the Community Innovation Survey, an unbalanced firm-level panel data set collected in the Netherlands between 1994 and 2004. This database allows us to span the entire innovation process from initial resources committed (R&D labor and the accumulated knowledge stock) to the final resulting sales volume of new products. We find that inefficiency accounts for between 50 and 92% of the unexplained between firm and over time variation in innovation output, with changes in efficiency explaining on average 62% of the between-firm variation in innovativeness. We do not find a significant difference in average inefficiency between those that do and those that do not cooperate with competitors. However, although government funding does not affect the marginal productivity of the knowledge stock and research labor, firms receiving government support are more efficient than those that do not. Finally, we find that more competitive firms are more innovative in terms of generating new product sales from innovations.

Keywords Innovation · Scale economies · Frontier

JEL D21 · G21 · L10 · O30

J.W.B. Bos (✉)
Maastricht University School of Business and Economics,
P.O. Box 616, 6200 MD Maastricht, The Netherlands
e-mail: j.bos@maastrichtuniversity.nl

R.C.R. van Lamoen
Risk Management Department, De Nederlandsche Bank,
Financial Markets Division, NL-1000 AB Amsterdam, The Netherlands
e-mail: r.c.r.lamoen@dnb.nl

M.W.J.L. Sanders
Utrecht School of Economics, Utrecht University,
3512 BL Utrecht, The Netherlands
e-mail: m.w.j.l.sanders@uu.nl

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_10

10.1 Introduction

The importance of technical change as a driving force of economic growth and prosperity has been widely recognized in the literature. Aghion and Howitt (1998, p. 151) state that "The chances of achieving sustainable growth depend critically on maintaining a steady flow of technological innovations." Empirically, Research and Development (R&D) was quickly established as a key input in generating these technological innovations. A seminal contribution in this literature was made by Griliches (1980), who finds that the productivity slowdown in US manufacturing during the period 1965–73 was largely due to the collapse of R&D investment. These results have inspired more studies confirming that R&D drives innovation at the firm level (Griliches 1986, 1998; Jaffe 1986), the industry level (Griliches and Lichtenberg 1984; Nadiri 1980) and across countries (Griliches and Mairesse 1983, 1991; Mansfield 1988).

Looking at the process of knowledge production, Mairesse and Mohnen (2002) defined *innovativeness* as the (unexplained) ability to turn innovation inputs into innovation output (analogous to (total factor) productivity in the production function). Innovativeness captures factors such as technological, organizational, cultural and environmental factors as well as waste and inefficiency (Mairesse and Mohnen 2002). Following the traditional growth accounting logic, the changes in innovation output can then be ascribed to changes in the innovation inputs and a residual that picks up changes in innovativeness. Efforts to understand this relationship between R&D efforts and innovative outputs have resulted in estimations of the knowledge production function (KPF), with increasing econometric rigor and increasing levels of detail and data quality (Coe and Helpman 1995; Park 1995; Engelbrecht 1997; Lichtenberg and Van Pottelsberghe de la Potterie 1998; Keller 2002; Guellec and Van Pottelsberghe de la Potterie 2004; Griffith et al. 2004).

From this literature we know that not every dollar or hour spent on R&D is equally well spent (Ulku 2004; Acs and Audretsch 1991). For example, the European 'innovation paradox' is based on the observation that, even after correcting for differences in R&D investments, countries in the European Union lag behind their US competitors in creating economic value from these investments (Figel 2006). And less spectacular, but possibly more important from an innovation management perspective, there are large differences in innovation output among firms in an industry (Thompson 2001; Cohen and Klepper 1996; Cohen and Levin 1989) even after controlling for R&D expenditures. In other words, there seems to be substantial heterogeneity in innovativeness.

The main contribution of this paper lies in accounting for and explaining this variation. In the empirical literature on the KPF, researchers to date still assume, usually implicitly, that all innovation takes place at the frontier and no waste of R&D inputs occurs. As Thompson (2001) puts it, there is only little evidence on differences in firm's ability to innovate because these can only be observed indirectly. We fill that gap by presenting an established method to estimate innovative ability. We show that innovativeness, like productivity, can conceptually be split

between (in)efficiency and technology (e.g., Weil 2008) . Moreover, not accounting for (in)efficiency (changes) and its determinants potentially biases the estimates of the parameters in the innovation function (Greene 2005).[1] Next, we empirically establish the differences between inefficiency and technology, using Stochastic Frontier Analysis (SFA), a method that is well-established in productivity analysis (Aigner et al. 1977; Battese and Corra 1977; Meeusen and van den Broeck 1977).

To the best of our knowledge, we are among the first to apply SFA to the estimation of the KPF. Wang (2007), Wang and Huang (2007) and Fu and Yang (2009) estimate the KPF at the country level and find that a substantial share of the cross country variation in innovativeness can be attributed to inefficiency. Gantumur and Stephan (2010) is the only (unpublished) paper we have found to apply this method to micro-data. They estimate a distance function and focus on entirely different research questions. They focus on the acquisition of external technology, while our study focuses on cooperation with competitors/other institutions and government funding. Their results show that the variance in inefficiency is about twice the variance of the remaining unexplained error in the German context.[2]

We estimate a knowledge production frontier using the Community Innovation Survey, an unbalanced firm-level panel data set collected in the Netherlands between 1994 and 2004. This database allows us to span the entire innovation process from initial resources committed (R&D labor and the accumulated knowledge stock) to the final resulting sales volume of new products.[3] Our analysis allows us to examine to what extent there are inefficiencies in the innovation processes. We find that indeed inefficiency accounts for between 50 and 92% of the unexplained between firm and over time variation in innovation output. Furthermore, changes in efficiency explain on average 62% of the between firm variation in innovativess as defined by Mairesse and Mohnen (2002). We find that larger firms are typically *less* efficient, but produce more innovative output per unit of input than small firms, suggesting that hierarchy and bureaucracy are bad for

[1]With the advent of endogenous growth theory (Romer 1990; Aghion and Howitt 1992; Grossman and Helpman 1991) precisely estimating unbiased parameters of the KPF has gained importance from a theoretical point of view. The scale effects in the first generation endogenous growth models (Jones 1995) depend on the value of the parameters in the innovation function. Sanders (2005) shows that the fate of the market size effect in Acemoglu (1998, 2002a, b) depends crucially on the degree of diminishing returns to labor in the innovation process and Ha and Howitt (2007) and Madsen (2008) estimated knowledge production functions to distinguish between second generation Schumpeterian (Aghion and Howitt 1998; Dinopoulos and Thompson 1998; Peretto 1998; Young 1998; Howitt 1999) and semi-endogenous growth models (Jones 1995; Kortum 1997; Segerstrom 1998).

[2]Several others have employed the closely related Data Envelopment Analysis (DEA) to estimate the innovation frontier (see Zhang et al. 2003). Coelli et al. (2005) provides a discussion of the differences between DEA and SFA.

[3]We excluded process innovations because typically one firm's process innovation is another (upstream) firm's product innovation.

innovative efficiency but do not overturn the positive effects of scale economies in R&D.[4]

Our data also allow us to explore the impact of cooperative innovation activities and government support on the productivity of innovation inputs and inefficiency in the innovation process. We do not find a significant difference in average inefficiency between those that do and those that do not cooperate with competitors. However, firms receiving government support are more efficient than those that do not. Whether this is due to a selection effect or because subsidies enable firms to learn to be more efficient is something we cannot establish with certainty given the short length of our panel. The innovativity (marginal productivity) of the knowledge stock and research labor does not differ between firms with and without cooperative innovation activities or between firms with and without government funding.

Finally, by combining our data with information from the Dutch Production Statistics we can also relate the estimated (in)efficiencies in the innovation process to firms' price-cost margin. Interpreting this margin as a measure of competition, we find that more competitive firms are more innovative in terms of generating new product sales from innovations.

The remainder of the paper is structured as follows. Section 10.2 describes the methodology used and data collected to estimate the innovation function. In Sect. 10.2, we describe our results. Section 10.4 concludes.

10.2 Methodology and Data

10.2.1 Methodology

We follow endogenous growth theory and assume that firms produce innovations using an accumulated knowledge stock in combination with a flow of R&D labor (Jones 1995; Pakes and Schankerman 1984). Keeping the stock of knowledge constant we can draw the knowledge production function as an innovation frontier, as in Fig. 10.1.

The innovation output function is concave if we assume diminishing returns to R&D labor.[5] In this figure we have drawn three hypothetical observations in our data. Firms (*a*), (*b*) and (*c*) are subject to the same innovation frontier, share the same knowledge stock and still operate at different innovative output levels. Since firms (*a*) and (*c*) operate *on* the innovation frontier without inefficiencies in their innovation process, firm (*c*) has less innovation output than firm (*a*) simply because

[4]This result overturns the finding in Gantumur and Stephan (2010), who find that larger firms tend to be more innovative. They did not, however, control for firm size in the innovation production frontier estimation.

[5]We do not impose diminishing returns in the empirical model below. The figure merely serves to illustrate the methodology.

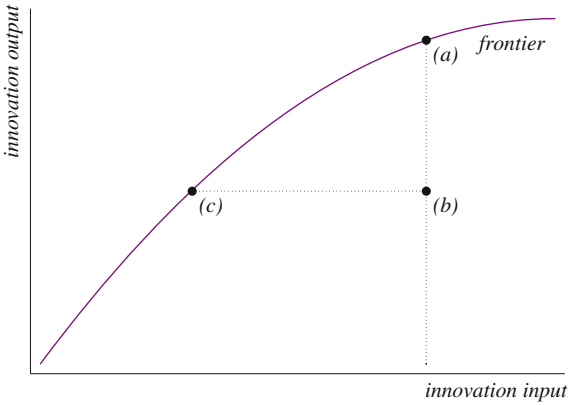


Fig. 10.1 Innovation frontier

it employs less R&D labor. Firm (*b*) produces less innovations than (*a*) despite the fact that they use the same amount of R&D labor. This discrepancy has to be attributed to the presence of inefficiencies in the innovation process of firm (*b*). As we cannot exclude a priori that inefficiency does not exist, we need to estimate the innovation frontier accounting for inefficiency. Put differently, we require an estimator that accommodates the fact that not all unexplained variance is pure noise, but some of it may capture the waste in the knowledge production process by firms such as firm (*b*).

The estimator that we use is rooted in stochastic frontier analysis, or SFA. First introduced by Aigner et al. (1977), Battese and Corra (1977) and Meeusen and van den Broeck (1977), SFA has been developed and applied in productivity analysis at the micro and macro-level (Kumbhakar and Lovell 2000). The innovation frontier defines the maximum innovative output achievable, given the current production technology and available inputs. If all firms produce on the boundary of a common knowledge production set that consists of an input vector with two arguments, the accumulated knowledge stock (A) and R&D labor (R), innovative output X of each firm can be described as:

$$X_{it}^{*} = f(A_{it}, R_{it}, t; \beta) \exp\{v_{it}\}, \tag{10.1}$$

where X_{it}^{*} is the firm's frontier (optimum) level of innovative output; f and parameter vector β characterizes the production technology; t is a time trend variable that captures neutral technical change (Solow 1957); and v_{it} is and i.i.d. error term distributed as $N(0, \sigma_v^2)$, which reflects the stochastic nature of the frontier.

In the presence of possible inefficiencies, we can express the difference between the optimum and actual (observable) innovative output by an exponential factor, $\exp\{-u_{it}\}$. In that case, we can express the actual innovative output, X_{it} as $X_{it} = X^*_{it} \exp\{-u_{it}\}$, or equivalently:

$$X_{it} = f(A_{it}, R_{it}, t; \beta) \exp\{-u_{it}\} \exp\{v_{it}\}, \tag{10.2}$$

where $u_{it} \geq 0$ is assumed to be i.i.d., with a normal distribution truncated at μ, $\left|N(\mu, \sigma_u^2)\right|$ and independent from the noise term, v_{it}.[6]

To also allow for flexibility in the functional form we estimate the frontier as a translog innovation function:

$$\begin{aligned} \ln X_{it} = {} & \beta_A \ln A_{it} + \beta_R \ln R_{it} + \frac{1}{2}\beta_{AA} \ln A_{it}^2 + \frac{1}{2}\beta_{RR} \ln R_{it}^2 \\ & + \beta_{AR} \ln A_{it} \ln R_{it} + \tau_t D_t + \beta_j + \beta_i + v_{it} - u_{it}, \end{aligned} \tag{10.3}$$

where D_t are time dummies that capture 'technical change' in the innovation process, β_j are technology class dummies and β_i are firm specific random effects over time. By including time dummies in the specification, we allow for shifts in the innovation frontier in a more flexible manner than using a time trend (Baltagi and Griffin 1988).[7] Dummies are also included to control for the industry technology class to allow for different innovation frontiers in e.g. pharmaceuticals, electronics and steel. These dummies are based on the OECD industry classification.[8]

In addition we want to control for firm size and the intensity of product market competition. The reason to control for size is that large and small firms report differently on their R&D inputs (Kleinknecht et al. 1991) and this systematic measurement error might affect our estimations. The control for competition intensity is necessary because our proxy for innovation output, innovative product sales, may be systematically higher for firms in less competitive markets by the fact that market power allows them to charge higher prices for the same innovation output. We also include dummies for cooperation with competitors and other institutions, for funding from government agencies. Our full model specification is then given by:

[6]When estimating Eq. (10.2), we obtain the composite residual $\exp\{\upsilon_{it}\} = \exp\{-u_{it}\}\exp\{v_{it}\}$. Its components, $\exp\{-u_{it}\}$ and $\exp\{v_{it}\}$, are identified by the $\lambda = (\sigma_u/\sigma_v)$ for which the likelihood is maximized (for an overview, see Coelli et al. 2005).

[7]In contrast to studies that estimate a normal production function, when estimating the KPF there is no reason to assume a constant time trend. There is no reason to expect that "technological change" makes knowledge production more or less innovative over time in the same way that new technologies improve productivity in the production process.

[8]See Raymond et al. (2009), Appendix A.

$$\begin{aligned}\ln X_{it} = {} & \beta_A \ln A_{it} + \beta_R \ln R_{it} + \frac{1}{2}\beta_{AA} \ln A_{it}^2 + \frac{1}{2}\beta_{RR} \ln R_{it}^2 \\ & + \beta_{AR} \ln A_{it} \ln R_{it} + \tau_t D_t + \beta_z z_{it} + \beta_{Az} \ln A_{it} z_{it} \\ & + \beta_{Rz} \ln R_{it} z_{it} + \frac{1}{2}\beta_{AAz} \ln A_{it}^2 z_{it} + \frac{1}{2}\beta_{RRz} \ln R_{it}^2 z_{it} \\ & + \beta_{ARz} \ln A_{it} \ln R_{it} z_{it} + \beta_C C_{it} + \beta_{FS} FS_{it} + \beta_j \\ & + \beta_i + v_{it} - u_{it},\end{aligned} \tag{10.4}$$

where z_{it} is the vector of dummy variables that indicates whether firms cooperate or not and whether they receive funding or not, C_{it} is our measure of competition intensity and FS_{it} represents firm size.

After obtaining the estimated parameters of the innovation frontier we can compute the (in)efficiency for each firm. Their efficiency scores are measured as the ratio of actual over the maximum attainable innovation output that firms would have on the frontier, where $0 \leq \exp\{-u_{it}\} \leq 1$, and $\exp\{-u_{it}\} = 1$ implies full efficiency. In a second equation we related inefficiency u_{it} in the stochastic frontier model to our dummy variables and controls to see if (in)efficiency responds to these variables. Technical inefficiency u_{it} in our model is specified as:

$$u_{it} = \gamma_z z_{it} + \gamma_C C_{it} + \gamma_{FS} FS_{it} + w_{it}, \tag{10.5}$$

where the noise term w_{it} is defined by the truncation of the normal distribution with zero mean and variance σ_w^2. We use a one-step model, where the specified stochastic frontier model in Eq. (10.4) and the endogenous inefficiency term in Eq. (10.5) are estimated in a single step by maximum likelihood.[9] The estimated coefficients, γ_z, γ_C and γ_{FS}, now relate cooperation, funding, competition and firm size to the efficiency of the innovative process, i.e. the distance between firms (*a*) and (*b*) in Fig. 10.1 is related to our variables of interest. Cooperation, government funding, competition and firm size are thus allowed to affect both the position and shape of the frontier and the distance to the frontier.

Mairesse and Mohnen (2002) emphasize the importance of what remains to be explained in the production of innovations (innovativeness or TFP). Based on the basic specifications, we also decompose productivity change (innovativeness or TFP) by identifying (the share of) pure technical change, a scale component (economies of scale in employing a larger knowledge stock and more labor) and efficiency change to analyse the importance of inefficiency in the innovation production process. Finally, the extended specifications, including interaction terms between the innovation inputs and dummy variables that represent cooperative innovation activities and funding from the government, are then presented to

[9]In two-step estimations, stochastic frontier models are estimated first and the relationship between inefficiency and covariates in the second step. Wang (2002) show that two-step estimations produce biased estimates.

analyze whether cooperation with competitors, other institutions and funding from the government affect the innovativity and/or efficiency of the knowledge stock or labor in the innovation production process. However, we first describe our data and measures before turning to the results.

10.2.2 Data

To estimate our innovation frontiers we use firm-level data on innovation from the Community Innovation Survey (CIS) in The Netherlands (Brouwer et al. 2008; Raymond et al. 2009). For thepurpose of this study, the CIS data are merged with financial information from the Production Survey (PS). The CIS data contain information on the R&D and other innovation activities of the firm, such as innovation expenditures, innovation activities conducted with other institutions, the effects of the innovation output (e.g. quality improvement, product differentiation etc.) and sources of the knowledge used to produce innovations. The PS data provide information on output, employment, value added, profit and other financial information. Both the CIS and PS data are collected by the *Centraal Bureau voor de Statistiek*. The sample from the CIS is based on five survey waves, namely CIS 2 (1994–1996), CIS 2.5 (1996–1998), CIS 3 (1998–2000), CIS 3.5 (2000–2002) and CIS 4 (2002–2004). In The Netherlands, each innovation survey is conducted every two years. The CIS and PS data are a combination of census data and a stratified random sample. The census data contain all firms with 50 employees or more and the stratified sample is based on firms with less than 50 employees. The stratum variables are the economic activity and the number of employees, where the economic activity of a firm is based on the Dutch standard industrial classification (SBI). Firms that are included in one survey only are excluded from the sample.[10] The population of interest are firms with at least 10 employees and positive sales.

In the CIS questionnaire, firms are asked first to provide general information on their economic activity, sales, number of employees etc. The second part of the questionnaire contains questions about the innovation activities of firms, such as their R&D activities, the percentage sales from new product/services, other innovation input expenditures, partnerships in innovation activities etc. Firms are asked to provide information on the second part of the CIS questionnaire if they affirm one of the three questions regarding: (1) whether firms developed new or strongly improved products (2) whether firms used new or strongly improved production processes (3) whether the firm has ongoing or abandoned innovation activities. Firms are classified as innovators if they affirm one of these three questions.

[10]We assume that attrition of the panel data occurs exogenously.

10.2.3 Innovation Output

We use the sales from new or improved products as our measure of innovation output, X_{it}.[11] The main advantage of this innovation output measure is that it captures innovations directly by measuring the introduction and the success of the newly developed products or services. Conventional innovation output measures such as patents or citation-weighted patents cannot capture the output of all innovative activity as many innovations are not patented and patented ideas are not always commercialised.[12] A drawback of the sales from new products as an innovation output measure is that firms may provide only rough estimates of their sales due to innovative products. This may induce measurement error in the regression analysis.[13] Another drawback of this measure is that the sales from new products may be influenced by the life cycle of a product.[14] In our analysis, firms with more than 50% of their sales from new products are therefore excluded from the sample.[15] Firms with only process innovations are also excluded since these firms have no sales from innovations new to the firm. Hence, our sample consists of firms with only product innovations or both product and process innovations.

10.2.4 Innovation Inputs

We follow Hall and Jones (1999) and construct a stock of total innovation expenditures by using a perpetual inventory method to proxy for A_{it}, the accumulated knowledge stock.[16] The stock of accumulated total innovation

[11]The analysis in this paper is restricted to products new to the firm instead of using an innovation output measure based on products new to the market. Brouwer et al. (2008) argue that a measure based on products new to the market may suffer from problems related to the interpretation of firms regarding their scope of the relevant market. This may lead to overestimation of innovation output by firms that are more focused on home markets.

[12]See Kamien and Schwartz (1982) and Geroski (1990) for a discussion on the limitations of patents as an innovation output measure.

[13]Measurement error in the dependent variable does not affect the consistency of the parameter estimates, however, if the component that represents the deviation from the true value of the dependent variable is not correlated with the (composite) error term or explanatory variables. We assume that this is the case here.

[14]New product sales follow a logistic curve as the product diffuses in the market. This implies that (small) firms that do R&D and introduce a new product may experience the large increases in their new product sales with quite a lag.

[15]This cut-off point is also used by Raymond et al. (2009).

[16]Hall and Jones (1999) use R&D expenditures instead of total innovation expenditures. The knowledge capital model has a well-known list of drawbacks (see e.g. Grilliches 2000) of which we are aware. By constructing it from total innovation expenditures some of these drawbacks have been addressed, but others remain. Not including a proxy for the knowledge stock, however, takes one far from what mainstream endogenous growth models assume.

expenditures represents the knowledge capital of a firm. While many papers do not account for learning effects in patent races or the innovation process, it is assumed in this study that the knowledge stock or accumulated innovation experience is a primary input in the innovation process. Doraszelski (2003) shows that firms have incentives to reduce R&D expenditures if their knowledge stock increases. R&D efforts in the past affect the probability to win an R&D race positively.

To construct our stock, we use total innovation expenditures because using R&D expenditures will understate the knowledge base of small firms (and consequently bias the output elasticity of the other inputs up). In addition to internal and external R&D expenditures, total innovation expenditures also include the purchase of rights and licenses to use external technology, and the purchase of advanced machinery and computer hardware devoted to the implementation of product and process innovations.[17]

We define the knowledge stock as:

$$A_{it} = (1-\delta)A_{it-1} + I_{it}, \tag{10.6}$$

where A_t is the knowledge stock and I_t represents the total innovation expenditures during period t. Furthermore, the depreciation rate δ is assumed to be 15% and the pre-sample growth rate of innovation expenditures g is 5%.[18] The knowledge stock at the beginning of the first period is defined by the following equation:

$$\begin{aligned} A_{i1} &= I_{i0} + (1-\delta)I_{i-1} + (1-\delta)^2 I_{i-2} + \ldots \\ &= \sum_{s=0}^{\infty} (1-\delta)^s I_{i-s} = I_{i0} \sum_{s=0}^{\infty} \left(\frac{1-\delta}{1+g}\right)^s = \frac{I_{i0}}{g+\delta}. \end{aligned} \tag{10.7}$$

In addition to the stock of knowledge, we included the amount of researchers in R&D activities in full-time equivalents as the second flow input, R_{it}, in the innovation process.

[17]Total innovation expenditures is only based on these components in our study. The total innovation expenditures in the pre-CIS4 data also includes additional components. However, these components are not included as measures in CIS4 and are excluded to ascertain the consistency of this measure across CIS waves.

[18]These values for the depreciation rate and growth rate are often used to construct the knowledge stock (Hall and Mairesse 1995). The results are robust when a 10 and 20% depreciation rate for the knowledge stock are used.

10.2.5 Competition, Firm Size, Cooperation and Funding

Firm size (FS_{it}) is measured by the total number of employees in the firm. Two cooperation dummy variables indicate whether firms cooperate with competitors (COOPCOMP) and other institutions (COOPOTHER) in their innovation activities. We also include a dummy variable to examine the effect of R&D funding from the government (FUNDING), where the value 1 indicates that the firm received funding from the local government, national government or European Union. The reference group for this dummy variable consists of firms without R&D funding from the government. Competition intensity cannot be observed directly, but proxies have been suggested in the literature. As in Aghion et al. (2005), we use the price cost margin (viz., the Lerner index, or markup) as our measure of competition. The price cost margin is calculated by dividing the total sales minus the cost of sales, e.g. labor expenses and energy costs, by total sales:

$$C_{it} = \left(\frac{S_{it} - TVC_{it}}{S_{it}}\right), \tag{10.8}$$

where C_{it} is the competition variable, S_{it} total sales and TVC_{it} represents the total variable costs. We use a firm-level measure of competition, since industries are relatively broadly defined in the data set and the intensity of competition can differ between firms, even within narrowly defined industries.[19] Hence, we assume that all changes in competition are reflected in the price cost margins of firms. An important advantage of the price cost margin over conventional measures of competition such as the Herfindahl-Hirschman Index (HHI) and concentration ratios is that the price cost margin does not require a precise definition of the relevant geographical or product boundaries. To eliminate outliers in our measure we eliminated observations that fall outside the range −1 and 1.

Table 10.1 provides the descriptive statistics of innovation output, innovation inputs, cooperative innovation activities, funding from the government, the price cost margin and firm size. The means of the sales from innovations, the knowledge stock and research labor in fte are €7482; €3466 and 2.442 fte, respectively. Most of the firms in our data set are not cooperating on innovation activities. Only 11.8% of the firms cooperate with competitors on innovation activities, 37.8% cooperate with other institutions, while the remaining 60.6% of the firms are not cooperating on their innovation activities. More than half of the firms in the sample received funding from a local/regional authority, the central government or the European Union, namely 63.8%. On average, firms earn a price cost margin of 24.8% and have 187 employees on their payroll.

[19]Our sector classification is based on the 3-digit SBI.

Table 10.1 Descriptive statistics

Variable	Symbol	Mean	Std. dev.	Minimum	Maximum
Sales from innovations in €1000	X	7481.567	14,319.87	1.521	218,986.6
Knowledge stock in €1000	A	3466.236	4885.73	18.61	28,876.23
Research labor in fte	R	2.442	3.965	0.029	40
Cooperation with competitors	COOPCOMP	0.118	0.323	0	1
Cooperation with other institutions	COOPOTHER	0.378	0.485	0	1
Funding from the government	FUNDING	0.638	0.481	0	1
Price cost margin	C	0.248	0.1119	−0.816	0.704
Number of employees (own payroll)	FS	187.04	350.678	0	10,857

The descriptive statistics are based on the sample in Column iii in Table 10.2 (1366 observations)

10.3 The Results

We start our analysis with two basic specifications, that are relatively close to what has been estimated in the literature so far. Next, we present the results from our preferred specification, that is the most comprehensive, and captures as many aspects of the innovation process as is possible, given our data. We end with a brief robustness analysis, where we examine the importance of lagged effects.

10.3.1 Basic Specifications

Columns (i) and (ii) in Table 10.2 presents the results based on the basic specifications. The dependent variable is always the share of innovative sales new to the firm in total firm sales. To examine the presence of inefficiency in the innovation process, a likelihood-ratio test is performed assuming the null hypothesis of no technical inefficiency ($H_0 : \sigma_u = 0$). The null hypothesis is rejected at the 1% level and indicates the presence of inefficiency in the innovation process. In this basic specification, we find that (in)efficiency accounts for approximately 50% of the variation in the residual (the ratio of variation in (in)efficiency σ_u over total variation $\sigma_u + \sigma_v$).

Column (i) in Table 10.2 shows the results based on a translog innovation function, where inefficiency is related to cooperation with other competitors, cooperation with other institutions, funding from the government, competition and firm size. However, we assume that the innovation frontier is the same for all firms

Table 10.2 The innovation function

Specification	(i)	(ii)	(iii)	(iv)
Panel A: Basic parameters				
$\ln A_{it}$	0.176 (0.363)	−0.065 (0.267)	0.849* (0.459)	−0.578 (0.415)
$\ln R_{it}$	−0.957*** (0.296)	0.042 (0.221)	−0.595* (0.359)	0.129 (0.371)
$\frac{1}{2}\ln A_{it}^2$	−0.004 (0.049)	0.111* (0.058)	0.042 (0.036)	−0.089 (0.063)
$\frac{1}{2}\ln R_{it}^2$	−0.174*** (0.048)	−0.010 (0.035)	−0.018 (0.054)	−0.011 (0.068)
$\ln A_{it} \ln R_{it}$	0.178*** (0.039)	0.005 (0.049)	0.106** (0.029)	0.004 (0.047)
Panel B: Cooperation with competitors and the innovativity of A and R				
$COOPCOMP_{it}$		−0.020 (0.127)	−7.975** (3.865)	−1.980 (4.562)
$\ln A_{it} * COOPCOMP_{it}$			2.149** (1.007)	0.496 (1.227)
$\ln R_{it} * COOPCOMP_{it}$			−1.117 (0.796)	0.947 (0.952)
$\frac{1}{2}\ln A_{it}^2 * COOPCOMP_{it}$			−0.283** (0.130)	−0.070 (0.163)
$\frac{1}{2}\ln R_{it}^2 * COOPCOMP_{it}$			−0.143 (0.125)	0.297** (0.146)
$\ln A_{it} \ln R_{it} * COOPCOMP_{it}$			0.138 (0.102)	−0.153 (0.125)
Panel C: Cooperation with other institutions and the innovativity of A and R				
$COOPOTHER_{it}$		0.154* (0.086)	3.023 (2.522)	−0.635 (3.057)
$\ln A_{it} * COOPOTHER_{it}$			−0.789 (0.685)	0.496 (1.227)
$\ln R_{it} * COOPOTHER_{it}$			0.353 (0.577)	0.947 (0.952)
$\frac{1}{2}\ln A_{it}^2 * COOPOTHER_{it}$			0.103 (0.092)	−0.070 (0.163)
$\frac{1}{2}\ln R_{it}^2 * COOPOTHER_{it}$			−0.032 (0.084)	0.297** (0.146)
$\ln A_{it} \ln R_{it} * COOPOTHER_{it}$			−0.055 (0.076)	−0.145 (0.125)
Panel D: Funding from the government and the innovativity of A and R				
$FUNDING_{it}$		0.037 (0.037)	4.742** (1.965)	−0.852* (0.464)
$\ln A_{it} * FUNDING_{it}$			−1.387** (0.538)	0.098 (0.066)
$\ln R_{it} * FUNDING_{it}$			1.023** (0.444)	0.933** 0.372

(continued)

Table 10.2 (continued)

Specification	(i)	(ii)	(iii)	(iv)
$\frac{1}{2}\ln A_{it}^2 * FUNDING_{it}$			0.200*** (0.073)	−0.002 (0.003)
$\frac{1}{2}\ln R_{it}^2 * FUNDING_{it}$			0.076 (0.071)	0.173** (0.081)
$\ln A_{it} \ln R_{it} * FUNDING_{it}$			−0.153*** (0.059)	−0.142*** (0.049)
Panel E: Other controls that affect innovation output				
C_{it}		−1.361*** (0.342)	−1.332*** (0.340)	−1.512*** (0.416)
FS_{it}		0.005*** (0.001)	0.005*** (0.001)	0.006*** (0.001)
Panel F: Determinants of inefficiency				
$COOPCOMP_{it}$	−0.528*** (0.176)	0.155 (0.736)	−0.133 (0.723)	−1.043* (0.597)
$COOPOTHER_{it}$	−0.184** (0.091)	−0.238 (0.529)	−0.237 (0.522)	−0.283 (0.375)
$FUNDING_{it}$	−0.072 (0.082)	−0.727 (0.485)	−0.682 (0.472)	−0.912*** (0.341)
C_{it}	0.319 (0.420)	−3.405* (1.969)	−3.289* (1.876)	−2.344* (1.251)
FS_{it}	−0.003*** (0.0001)	0.006*** (0.001)	0.005*** (0.001)	0.006*** (0.001)
$\sigma_u/(\sigma_u + \sigma_v)$	0.499	0.918	0.920	0.825
Observations	1366	1366	1366	926
Technology class	Yes	Yes	Yes	Yes
Technical change	No	Yes	Yes	Yes

The dependent variable is sales from innovations. Standard errors (between parentheses) are robust against heteroskedasticity. The results in Column iv are based on a lagged effect of cooperation with competitors, other institutions and funding. Asterisks indicate significance at the following levels: *0.10, **0.05, and ***0.01

and do not allow for innovativity differences between firms with respect to the knowledge stock and labor. Only the research labor, its squared term and the interaction term are individually and jointly significant at the 1% level. The knowledge stock, its squared term and the interaction term are jointly insignificant

at the 1% level. The output elasticities with respect to the knowledge stock and labor are 0.16 and 0.33, respectively.[20]

The results in Panel F for Column (i) show the determinants of inefficiency. Cooperation with competitors (COOPCOMP) is significant at the 1% level and cooperation with other institutions (COOPOTHER) is significant at the 5% level. Both are negatively related to inefficiency and so cooperation reduces inefficiency, where cooperation with competitors is found to be more important for inefficiency compared to cooperation with other institutions. Funding from the government (FUNDING) and the price cost margins (C) are not significantly related to inefficiency in the production of innovations. Firm size (FS) is significantly negatively related to inefficiency at the 1% level. This means that larger firms are producing innovations more efficient (on average). This finding is consistent with the outcomes of Gantumur and Stephan (2010). Large firms may be more efficient if they use more specialised inputs in production. However, this specification assumes that all firms are subject to the same innovation frontier and does not allow for differences in the innovativity of the knowledge stock and labor between firms with and without cooperation on innovation activities or firms with and without government funding. Furthermore, the specification in Column (i) does not yet include firm size and competition in the innovation function directly.

Column (ii) therefore presents the results based on including cooperation with competitors, other institutions, competition and firm size directly in the innovation function and as explanatory variables for the inefficiency term.[21]

We observe an increase in the ratio of variance in (in)efficiency over total variance (to 92%), because introducing more controls in the specification of the frontier implies that the overall fit is improved with the same variance in (in)efficiency. Although the knowledge capital stock, research labor, quadratic terms and interaction term are individually insignificant, they are jointly significant at the 1% level. The output elasticities with respect to knowledge capital and labor are now 0.24 and 0.08, respectively.

In panels C and D, cooperation with competitors and funding from the government are not significantly related to innovation output directly. Cooperation with other institutions is positively related to the innovation output, but only at the 10% significance level. In Panel F, the results show that cooperation with competitors

[20]These output elasticities differ quite markedly from the output elasticities obtained when we exclude the determinants of inefficiency (not reported here). In that case, the marginal innovativity of the knowledge stock is much higher, while the marginal innovativity of research labor is lower. Furthermore, in that case the output elasticity of research labor is lower than the output elasticity of the knowledge stock. Compared the specification in Column (i), the variance attributed to (in)efficiency is then much higher: 99% compared to the current 50%.

[21]We also estimated this specification with interaction terms between the innovation inputs and technology classes to examine whether the marginal innovativity of the knowledge stock and research labor differ between technology classes (industry groups). The interaction terms are not jointly significant. This suggests that the marginal innovativity of the knowledge stock and research labor do not differ across technology classes.

and other institutions are no longer significantly related to inefficiency after the inclusion of these variables in the innovation function.

The price cost margin, however, is now significant at the 1% level and negatively related to innovation output in panel E. This means that more competition is positively related to the share of innovative sales. A possible explanation is that competitive pressure induces firms to introduce more innovations and weeds out unsuccessful ones faster.[22] Another explanation may be that new innovations cannibalise the sales from existing products and therefore lower the price cost margins. We only find a negative correlation and cannot infer the direction of causality in our data. Competition is also significant and negatively related to inefficiency at the 10% level in panel F. Firms with higher price cost margins thus seem to have lower inefficiency. This partially offsets the positive direct effect of competition on innovativeness found above.

Firms size remains significant at the 1% level in panel E but is now *positively* related to innovation output. An explanation for this result is that larger firms have developed larger distribution and marketing networks to sell their new products. In short, they can benefit from economies of scale in the innovation process. In contrast to the results in column (i), however, firm size is now also significantly *positively* related to inefficiency at the 1% level. Again we see that the benefits of economies of scale found above are partially offset by higher inefficiency in larger firms. Larger firms may suffer from coordination problems and therefore experience on average more inefficient innovation processes. These findings with respect to firm size are not in line with the findings of Gantumur and Stephan (2010) reported earlier. A likely explanation for the discrepancy between our findings and theirs is that firm size was not directly included in the innovation function in the specification used by Gantumur and Stephan (2010) as in our column ii.[23]

10.3.2 Preferred Specification

In column (iii) of Table 10.2 we show the results based on the interaction between the innovation inputs and the dummy variables that indicate whether firms cooperate with competitors or other institutions and whether they receive funding from the government. As in the specification in column (ii), the variance in inefficiency explains 92% of the total variance of the composite residual. F-tests are used to

[22]A positive relationship between price cost margins and the sales from innovations is expected if less competition allows firms to reap higher rents from their innovations.

[23]A drawback of the specification in Column iv is that the interaction between the innovativity of knowledge capital, labor and cooperation with institutions and funding from governments cannot be examined. These interactions are necessary to examine whether the marginal innovativity of the knowledge stock and labor differ between these groups. We have done these regressions and find no clear evidence of strong interaction effects. Therefore these results have been delegated and discussed in the appendix.

examine whether the innovativity (marginal productivity) of the knowledge stock and research labor differs between cooperating firms and firms that do not cooperate and firms with and without funding. The cooperation with competitors dummy and interaction terms with the knowledge stock and research labor are not jointly significant. Also the dummy variable that indicates cooperation with other institutions and interaction terms are not jointly significant. However, an F-test based on funding from the government and interaction terms indicates that there are significant differences in the innovativity of the knowledge stock and labor between firms with funding from the government and firms without funding.[24] The output elasticities of the knowledge stock for firms without funding and with funding are significant at the 1% level and 0.21 and 0.27, respectively.[25] The output elasticities of the research labor for firms without funding and with funding are significant at the 5% level and 0.18 and 0.09, respectively (see footnote 25). While the point estimate of the output elasticity with respect to the knowledge stock is somewhat higher for firms with funding from the government compared to firms without funding, the innovativity of research labor is lower for firms with funding. We also performed t-tests to examine whether the output elasticities differ significantly between firms with and without funding. There are no significant differences in innovativity between firms with and without funding based on the t-tests.[26]

When we examine the determinants of inefficiency we conclude that only competition and firm size are on average related to inefficiency. The relationship between competition and inefficiency in the innovation process is quite weak since it is only significant at the 10% level. Higher price cost margins (less competition) are related to lower inefficiency. Consistent with the findings in column iv, firm size is significantly positively related with inefficiency at the 1% level. Coordination problems in the production of innovations by large firms may lead to less efficient innovation production processes. Cooperation with competitors or other institutions and funding from governments are on average not related to inefficiency.[27]

[24]The dummy variable and interaction terms are significant at the 5% level.

[25]The output elasticities are evaluated at the average natural logarithm of the knowledge stock and research labor.

[26]The t-values with respect to the knowledge stock and research labor are 1.07 and 1.47, respectively.

[27]However, when we calculate efficiency scores and examine the differences in the distributions between groups, we find that the distributions of the efficiency scores differ between firms with and without funding from governments. A Kolmogorov-Smirnov test is used and the distributions differ significantly from each other at the 1% significance level. Based on a kernel density, we find that at higher efficiency scores, there are more firms with funding from governments than without funding. Furthermore, there are more firms without funding at lower levels of efficiency scores. Hence, firms with funding from the government are more likely to produce innovations efficiently compared to firms without funding from the government. We do not find differences in the distributions of efficiency scores between firms with and without cooperative innovation agreements.

10.3.3 Robustness Analysis

Column (iv) provides an overview of the estimation results based on a lagged effect of cooperative innovation agreements and government subsidies.[28] The variance of the inefficiency term accounts for most part of the total variance of the residual, namely 83%. With respect to cooperation on innovation activities, the F-tests indicate that the innovativity of the knowledge stock and research labor differs between firms that cooperate and firms that do not cooperate.[29] Furthermore, the tests show that there are differences between firms with funding from the government and firms without funding. However, evaluated at the average natural logarithm of the knowledge stock and research labor, we do not find that the point estimates of the output elasticities differ significantly between firms with and without cooperation on innovation and with and without government funding.[30] The price cost margin is significant at the 1% level and negatively related to the sales from innovations. This finding is consistent with the results in the previous specifications. Firm size is significant at the 1% level and larger firms have more sales from new innovations.

We also examined the lagged effect of cooperation with competitors, cooperation with other institutions and funding from the government on inefficiency in the innovation production process. Cooperation with competitors in the previous period is negatively related with inefficiency at the 10% level. Thus we find weak evidence that cooperation with competitors in the past leads to improved efficiency in the future. There is no significant relationship between cooperation with other institutions and inefficiency. Funding from the government is significantly related to inefficiency at the 1% significance level. Firms that received funding from the government in the previous period are on average more efficient in the current period. Competition is significant at the 10% level and as in earlier specifications, less competition is positively related to efficiency. Moreover, larger firms are significantly less efficient than smaller firms.

[28]The lagged effect corresponds to a delayed effect of two years since bi-annual data are used.

[29]Cooperation with competitors and the interaction terms are significant at the 5% level. Cooperation with other institutions and the interaction terms are significant at the 1% level.

[30]This conclusion is based on t-tests to examine whether each output elasticity differs from the output elasticity of firms without cooperation and without funding. The absolute t-values range between 0.05 and 1.15. Nevertheless, differences in output elasticities may also arise due to different levels of the knowledge stock and labor. The change in the knowledge stock differs significantly at the 5% level between firms that received funding only in the previous period compared to firms that received no funding. The average change in the knowledge stock is 462,000 € for firms with funding and 76,000 € for firms without funding. However, we do not find significant differences in the output elasticities between these types of firms evaluated at their average knowledge stock.

10.4 Conclusion

This paper contributes to the innovation literature by examining several sources of innovativity and efficiency in the innovation process. We use Stochastic Frontier Analysis and estimate innovation frontiers for Dutch firms over the period 1994–2004. Dutch Community Innovation Survey data is used to examine whether cooperation with competitors, cooperation with other institutions, funding from the government, competition and firm size affect innovativity and efficiency in the production of innovations. We find that inefficiency is present in the innovation process of Dutch firms and that percentage changes in efficiency contributes on average 63% to the drop in innovativity that was observed in our sample. In addition, in our preferred specification, inefficiency explains 92% of the unexplained between firm variation in innovation outputs. Clearly inefficiencies are to be reckoned with in estimating the knowledge production function to avoid biased parameter estimates and investigate the sources of inefficiency.

On that last note we also find that cooperation with competitors and funding from the government are on average significantly negatively related to inefficiency in the basic specification where the innovation frontier is assumed to be homogenous for all firms. There is no relationship on average between cooperation with competitors, funding from the government and inefficiency, however, once we allow these factors to affect innovation output directly. In our preferred specification, we also found that competition is significantly positively related to the level of innovation output. Finally, our results also suggest that while larger firms have higher levels of innovation output, their innovation processes seem to be subject to more inefficiencies. Such conclusions could have important managerial and policy implications. Future research should therefore aim at both theoretically and empirically examining the channel through which firm size and competition affect innovativity and inefficiency in the innovation production process.

Acknowledgements We thank Jacques Mairesse and Pierre Mohnen for helpful discussions. The usual disclaimer applies.

References

Acemoglu D (1998) Why do technologies complement skills? Directed technical change and wage inequality. Q J Econ 113:1055–1090

Acemoglu D (2002a) Directed technical change. Rev Econ Stud 69(4):781–809

Acemoglu D (2002b) Technical change, inequality and the labor market. J Econ Lit 40:7–72

Acs ZJ, Audretsch DB (1991) R&D, firm size, and innovative activity. In: Acs ZJ, Audretsch DB (eds) Innovation and technological change: an international comparison. University of Michigan Press, Ann Arbor, pp 39–59

Aghion P, Howitt P (1992) A model of growth through creative destruction. Econometrica 60 (2):323–351

Aghion P, Howitt P (1998) Endogenous growth theory. MIT Press, Massachusetts

Aghion P, Bloom N, Blundell R, Griffith R, Howitt P (2005) Competition and innovation: an inverted-U relationship. Quart J Econ 120(2):701–728
Aigner DJ, Lovell KC, Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. J Econ 6(1):21–37
Baltagi BH, Griffin JM (1988) A general index of technical change. J Polit Econ 96(1):20–41
Battese GE, Corra GS (1977) Estimation of a production frontier model, with application to the pastoral zone of eastern Australia. Aust J Agric Econ 21(3):169–179
Brouwer E, Poot T, Van Montfort K (2008) The innovation threshold. De Econ 156:45–71
Coe DT, Helpman E (1995) International R&D spillovers. Eur Econ Rev 39(5):859–887
Coelli T, Rao DP, Battese GE (2005) An introduction to efficiency analysis, 2nd edn. Springer, New York
Cohen WM, Klepper S (1996) A reprise of size and R&D. Econ J 106(437):925–951
Cohen WH, Levin R (1989) Empirical studies of innovation and market structure. In: Schmalensee R, Willig R (eds) The handbook of industrial organization. North-Holland, pp 1060–1107
Dinopoulos E, Thompson P (1998) Schumpeterian growth without scale effects. J Econ Growth 3:313–335. doi:10.1023/A:1009711822294
Doraszelski U (2003) An R&D race with knowledge accumulation. RAND J Econ 34(1):20–42
Engelbrecht H-J (1997) International R&D spillovers, human capital and productivity in OECD economies: an empirical investigation. Eur Econ Rev 41(8):1479–1488
Figel J (2006) The European response on the innovation paradox. In: Conference "Innovation paradox: the Flemish response? the EU response", 21 Nov 2006
Fu X, Yang QG (2009) Exploring the cross-country gap in patenting: a stochastic frontier approach. Res Policy 38(7):1203–1213
Gantumur T, Stephan A (2010) Do external technology acquisitions matter for innovative efficiency and productivity? Technical Report 222
Geroski PA (1990) Innovation, technological opportunity, and market structure. Oxford Econ Papers 42(3):586–602
Greene WH (2005) Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. J Econ 126(2):269–303
Griffith R, Redding S, Reenen JV (2004) Mapping the two faces of R&D: productivity growth in a panel of OECD industries. Rev Econ Stat 86(4):883–895
Griliches Z (1980) R&D and the productivity slowdown. National Bureau of economic research working paper series no. 434
Griliches Z (1986) Productivity, R&D, and the basic research at the firm level in the 1970s. Am Econ Rev 76(1):141–154
Griliches Z (1998) Productivity and R&D at the firm level. In: R&D and productivity: the econometric evidence. NBER, pp 100–133
Grilliches Z (2000) R&D, education and productivity: a retrospective. Harvard University Press, Cambridge
Griliches Z, Mairesse J (1983) Comparing productivity growth: an exploration of French and U.S. industrial and firm data. Eur Econ Rev 21(1–2):89–119
Griliches Z, Lichtenberg F (1984) R&D and productivity growth at the industry level: is there still a relationship? In: Griliches Z (ed) R&D, patents and productivity. University of Chicago Press, Chicago
Griliches Z, Mairesse J (1991) R&D and productivity growth: comparing japanese and U.S. manufacturing firms. Technical Report 1778
Grossman G, Helpman E (1991) Innovation and growth in the global economy. MIT Press, Massachusetts
Guellec D, Van Pottelsberghe de la Potterie B (2004) From R&D to productivity growth: do the institutional settings and the source of funds of R&D matter? Oxford Bull Econ Stat 66(3):353–378

Ha J, Howitt P (2007) Accounting for trends in productivity and R&D: a schumpeterian critique of semi-endogenous growth theory. J Money Credit Banking 39(4):733–774
Hall BH, Mairesse J (1995) Exploring the relationship between R&D and productivity in French manufacturing firms. J Econ 65(1):263–293
Hall RE, Jones CI (1999) Why do some countries produce so much more output per worker than others? Quart J Econ 114(1):83–116
Howitt P (1999) Steady endogenous growth with population and R&D inputs growing. J Polit Econ 107(4):715–730
Jaffe AB (1986) Technological opportunity and spillovers of R&D: evidence from firms' patents, profits, and market value. Am Econ Rev 76(5):984–1001
Jones CI (1995) R&D-based models of economic growth. J Polit Econ 103(4):759–784
Kamien MI, Schwartz NL (1982) Market structure and innovation. Cambridge University Press, New York
Keller W (2002) Geographic localization and international technology diffusion. Am Econ Rev 92 (1):120–142
Kleinknecht A, Poot TP, Reijnen JON (1991) Formal and informal R&D and firm size. Survey results from The Netherlands. In: Acs ZJ, Audretsch DB (eds) Innovation and technological change: an international comparison. University of Michigan Press, Ann Arbor, pp 84–108
Kortum SS (1997) Research, patenting, and technological change. Econometrica 65(6):1389–1419
Kumbhakar SC, Lovell KC (2000) Stochastic frontier analysis. Cambridge University Press, Cambridge
Lichtenberg FR, Van Pottelsberghe de la Potterie B (1998) International R&D spillovers: a comment. Eur Econ Rev 42(8):1483–1491
Madsen J (2008) Semi-endogenous versus schumpeterian growth models: testing the knowledge production function using international data. J Econ Growth 13:1–26
Mairesse J, Mohnen P (2002) Accounting for innovation and measuring innovativeness: an illustrative framework and an application. Am Econ Rev 92(2):226–230
Mansfield E (1988) Industrial R&D in Japan and the United States: a comparative study. Am Econ Rev 78(2):223–228
Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb-Douglas production functions with composed error. Int Econ Rev 18(2):435–444
Nadiri MI (1980) Sectoral productivity slowdown. Am Econ Rev 70(2):349–352
Pakes A, Schankerman M (1984) R&D, patents and productivity. NBER, Ch. The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources, pp 73–88
Park WG (1995) International R&D spillovers and OECD economic growth. Econ Inq 33(4):571–591
Peretto PF (1998) Technological change and population growth. J Econ Growth 3(4):283–311
Raymond W, Mohnen P, Palm F, Schim van der Loeff S (2009) Innovative sales, R&D and total innovation expenditures: panel evidence on their dynamics. Technical Report 028
Romer PM (1990) Endogenous technological change. J Polit Econ 98(5):71–102
Sanders M (2005) Market size and acceleration effects; comparing hypotheses to explain skill biased technical change. In: Max Planck Institute of Economics Working Papers 2005-003
Segerstrom PS (1998) Endogenous growth without scale effects. Am Econ Rev 88(5):1290–1310
Solow RM (1957) Technical change and the aggregate production function. Rev Econ Stat 39 (3):312–320
Thompson P (2001) The microeconomics of an R&D-based model of endogenous growth. J Econ Growth 6(4):263–283
Ulku H (2004) R&D, innovation, and economic growth: an empirical analysis. Technical Report 04/185
Wang EC (2002) Public infrastructure and economic growth: a new approach applied to East Asian economies. J Policy Model 24(5):411–435

Wang EC (2007) R&D efficiency and economic performance: a cross-country analysis using the stochastic frontier approach. J Policy Model 29(2):345–360
Wang EC, Huang W (2007) Relative efficiency of R&D activities: a cross-country study accounting for environmental factors in the DEA approach. Res Policy 36(2):260–273
Weil DN (2008) Economic growth, International edn. Pearson Education, New York
Young A (1998) Growth without scale effects. J Polit Econ 106(1):41–63
Zhang A, Zhang Y, Zhao R (2003) A study of the R&D efficiency and productivity of Chinese firms. J Comp Econ 31(3):444–464

Chapter 11
Potential Coopetition and Productivity Among European Automobile Plants

Jonathan Calleja-Blanco and Emili Grifell-Tatjé

Abstract The chapter proposes a definition of the potential economic incentives for competitors to cooperate with each other, namely *coopetition*. A non-parametric methodological approach based on the rate of return on assets (ROA), a well-known measure of financial performance, enables comparison between non-coopetition and coopetition statuses. The potential ROA gains from competition are decomposed by economic drivers. This methodology was applied to the study at plant level, focusing on cases of potential competition in the European automotive industry. The main results are based on an analysis of a generated sample of over forty-five thousand cases of potential cooperation between plants in the 2000–2012 period. Out of that sample, roughly twelve thousand cases (about 27%) presented potential ROA gains from coopetition. Results show that faster asset turnover and better productivity explain a higher potential ROA from coopetition. Results also reveal that medium–small and small plants have the strongest economic incentive for coopetition. The chapter concludes by offering some policy recommendations concerning the introduction of changes to the legal framework of competition, in the context of the European Union.

Keywords Productivity · Coopetition · Plant level · Automobile industry · Return on assets (ROA) · DEA

11.1 Introduction

—The auto industry of the future is collaborative and borderless

"Automotive 2025: Industry without borders" IBM report

J. Calleja-Blanco · E. Grifell-Tatjé (✉)
Department of Business, Universitat Autònoma de Barcelona, Barcelona, Spain
e-mail: emili.grifell@uab.cat

J. Calleja-Blanco
e-mail: j.calleja.blanco@gmail.com

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_11

It is no longer enough to play individually in the automobile industry in order to be competitive. Separate efforts aimed at improving performance are paying off less and less. As a result, companies are looking for new ways to supply their customers while staying competitive by responding to new market trends, e.g. customization. The main producers, while being competitors, have identified the opportunities that cooperation offers. Firms look for suitable partners, including rivals, in order to enhance competitiveness. This study is a novel attempt to analyze, from an economic perspective, the potential or a priori impact of cooperation among independent automobile production plants in Europe from 2000 to 2012. In other words, we analyze the potential economic prospects as a basis for managerial decisions regarding cooperation.

The literature has devoted attention to the benefits of potential mergers, which are more likely to occur at firm level (e.g. Bogetoft and Wang 2005; Bagdadioglu et al. 2007; Kristensen et al. 2010; Halkos and Tzeremes 2013; Zschille 2015). In contrast, the economic gains that could be potentially achieved when cooperating with a competitor have been under-explored. The chapter contributes in this regard and can be placed within a *coopetition* framework: cooperation amongst competitors, i.e. the situation in which organizations that would normally compete with each other engage in a cooperative strategy to develop joint production. Hence, cooperation and competition, which are frequently studied separately, come together as part of the same strategy.

It is important to stress the difference between coopetition and collusion. Both strategies need cooperation, reducing overall competitiveness. Coopetition has the potential for collusive behavior and sometimes they are treated equally. However, the difference lies in the effect on the consumer. While collusion generally occurs in downstream activities, typically agreeing in prices, coopetition refers to upstream ones (Walley 2007; Rusko 2011). Under coopetition, firms can still compete in downstream activities.

From an engineering perspective, there have been some efforts to approach the idea of coopetition. Based on the virtues of 'commonizing' technologies to produce similar products, some authors have analyzed the implications of such a strategy (Muffatto 1999; Pasche and Sköld 2012). This step, technically speaking, paves the way towards cooperation with a counterpart in order to take advantage of well-known technologies, equipment and machinery. This cooperation based on common technologies can optimize the cost of coordination between independent plants. A plant might cooperate with some products and not cooperate with others. Of course, competition in the market exists regardless of the decision about coopetition. There is competition in the products that are not the object of coopetition and for end products when coopetition occurs in the production of intermediate goods. It is also possible regarding products that are the direct result of coopetition, which may be commercialized under different channels and brands.[1]

[1]An illustrative example of cooperation and competition outside the automobile industry is shown for Sony and Samsung (Gnyawali and Park 2011: 655).

Platform sharing has been a natural response from automobile manufacturers to improve performance, making production plants more flexible. General Motors first did this during the inter-war period (Freyssenet and Lung 2007) and many other actors have followed the same strategy since. Nowadays, almost all producers are trying to reduce the number of platforms.[2] However, platform-sharing among independent plants remains less common in Europe.

There are two possible explanations to consider why the degree of coopetition among European plants is low. The first is purely economic. There must be an economic incentive in order for plants to be willing to engage in a coopetitive strategy. The final economic outcome of cooperation lies beyond the scope of the study (which includes trust, commitment and bargaining agreements), but an analysis of potentiality is offered, for actors to be able to identify their best options. In tandem with the economic reason, there are legal limitations that cause plants not to cooperate more. We discuss these limitations in the conclusion section. This chapter concludes that there are potential economic incentives for coopetition between European automobile plants. Hence, the legislation on competitiveness issued from the European Union (EU) and its member states must partially justify the poor level of coopetition.

This chapter contributes to this area in several aspects. The potential economic benefits that separate agents might obtain when they commit to coopetition are quantified. A new methodology to define the economic incentives for coopetition is presented in Sect. 11.3. It contributes to the need expressed in the literature with regard to further exploration of coopetition outcomes (Gnyawali and Park 2011), as well as partner selection tools (Alves and Meneses 2013). Furthermore, Bouncken et al. (2015) also consider efficiency as one of the potential dimensions of coopetition that urgently needs to be developed. Additionally, Blum (2009) discusses the need for more research into the quantification of the potential economic gains associated with coopetition. Section 11.3 can be seen as a response to the need for more research, from an economic perspective, and exploits the latest findings by Grifell-Tatjé and Lovell (2015), who have introduced productivity as one of the drivers of return on assets change. It is shown that this approach naturally accommodates Bogetoft and Wang (2005), who have been mainly used in the literature to study potential mergers (Kristensen et al. 2010). The methodology developed in Sect. 11.3 may be of interest to practitioners. Its application enables the identification of the best potential options for coopetition. Before that, Sect. 11.2 presents the background of the European automobile industry. An introduction to the applied part of the chapter is given in Sect. 11.4, where the

[2]It is important to clarify that a platform is a construction system (a sort of architecture that defines the main design, engineering, etc. of a vehicle). A producer may have different production plants and still use a single type of platform to produce several car models (e.g. 'Ford plans to trim global vehicle platforms from 15 to 9 by 2016'). Industry experts expect almost half of world production to be manufactured using 20 core platforms in the coming years.

dataset is discussed. Section 11.5 presents the main results based on the study of a generated sample of over forty-five thousand cases of potential cooperation. Section 11.6 provides a set of conclusions, which have implications for industrial policy, the main point being the need to revise the European law on competitiveness in the manufacturing industry. Particular attention might be paid to automobile industry, for which legal framework is especially rigid.

11.2 Background

11.2.1 Previous Research on Coopetition

Although coopetition has become an accepted term in the management literature as a suitable firm strategy, the body of empirical articles studying the phenomenon still lacks a common definition. It sometimes overlaps with the idea of alliance, or the two notions are taken as part of the same action. A broad concept defines coopetition as a business relationship in which firms cooperate between and compete against each other simultaneously (Bengtsson and Kock 2000). This characterization is still broad in its scope and allows for many different configurations. Thus, empirical studies have made use of many singular boundaries of the concept, while some revisionist literature has pointed out the need for a more refined terminology (Bengtsson and Kock 2014; Bouncken et al. 2015). Cooperation among competitors has been analyzed without framing it under the name of coopetition (Oliver 2004).

A central aspect associated with successful coopetition, of any kind, is the will among managers for cooperation. The ability of partners to strike a balance between cooperation and competition determines success and also requires a new orientation of management (Peng and Bourne 2009). Some studies have gone in this direction by proposing guidelines for managers to achieve successful coopetition. Based on a literature review, Chin et al. (2008) rank commitment, relationship development and communication as the key factors in order for a partnership to work.

Bengtsson and Kock (2000) dissect the definition of coopetition according to the proximity of the end client. Cooperation is generally far from the end client and competition occurs at a closer stage to the same, so that each part might be managed differently. Strategies are, in that order, related to value creation and value capturing, and may be of different relative importance in the agreement (Luo 2007). Also, a greater number of similarities between products and technologies causes greater cooperation. Many firms cooperate at the initial stages of the product, preserving the competitive advantage for the final customization or sale. Wolff (2009), borrowing the term from a manager in the car industry, defines this situation as a *pre-competitive stage*, meaning cooperation in the generation of somewhat similar outcomes.

In the context of this pre-competitive state, many studies have focused on the benefits for innovativeness when engaging in coopetition, i.e. the initial stages of product development. Empirical results have proved the positive impact of coopetition on innovation (Li et al. 2011; Bouncken and Fredrich 2012; Ritala and Sainio 2014), knowledge creation (Zhang et al. 2010) and co-creation or technology development (Wilhelm and Kohlbacher 2011, in the Toyota network). In general, this stream has found some type of value creation based on innovation (Ritala and Hurmelinna-Laukkanen 2009).

While innovation development may imply some sort of mutual investments, cooperating solely in production would only need a certain type of input complementarities (Biesebroeck 2007). In other words, what is needed is the correct reallocation of existing complementary resources. In fact, this is horizontal cooperation based on redundant capabilities. These capabilities and competitive market forces are the main factors dragging firms to cooperate (Madhok 1996). Hence, coopetition in production must be based on the sharing of resources and technology up to a pre-competitive state. To our knowledge, the literature has paid little attention to this kind of coopetition. One exception is Ehrenmann and Reiss (2011), who advocate manufacturing firms to build up coopetition, in order to achieve their full performance potential. Here, excess capacity and mass customization are particularly important for the case of the automobile industry. This kind of coopetition, which is mainly based on the reallocation of existing complementary resources, should deliver higher productivity and output quantities. This is the main object of study in Sect. 11.3 of this chapter.

The next section depicts some examples of coopetition, from a somewhat broad perspective, that have appeared in the automobile industry, especially in Europe. These examples are mainly on a firm level given the scarce literature on the plant level, which is our unit of analysis.

11.2.2 Coopetition in the Automobile Industry

Car platforms have become a common practice in the automotive industry since General Motors initiated the concept. Automakers use platform sharing to combine lower-volume customized production with higher-volume standardized production. Thus, by sharing common technologies among different products, they are able to develop an additional number of models. It was initially designed for building cars of the same brand, or for cars belonging to the same matrix. Nowadays, competing groups are also integrating this tactic to share production with each other. It is an alternative to the wave of mergers that appeared a few decades ago. What is clear from the observance of this tendency is that remaining independent in the modern-day car industry is not only difficult but also inefficient.

An early case of this type of cooperation was the Portuguese Autoeuropa plant, settled as a 50/50 joint-venture between Volkswagen and Ford. Set up in 1991, this was an important player in the European production for both matrixes. For years, the cooperation paid off for both participants, but this sort of agreement is flexible enough to allow for exit when the initial interests disappear. Accordingly, Ford left the venture in 1999, and the German group now fully owns and manages the plant.

More recently, Volkswagen developed a car platform together with Ford's European subsidiary, which has mainly been used by the former to produce several of its cars. Daimler has also used Volkswagen-based technologies to produce some of its models, which they both commercialize in different markets. Daimler and Renault, Toyota and General Motors, Peugeot-Citroën and BMW, or General Motors and Peugeot-Citroën are other examples of established or potential (currently at the initial agreement stage) collaborations in the industry. Nevertheless, it is also true that some have ended wrongly, such as General Motors and Fiat not achieving successful results, or Renault and BMW withdrawing at the initial agreement stage.

By mid-2015, Toyota and Mazda had announced an agreement to create a partnership aimed at sharing the production of future car models. In their words, this partnership would "go beyond the traditional framework of cooperation, aiming instead to create a whole new set of values for cars through wide-ranging medium to long term collaboration". However, the pact still does not affect their individuality, as they keep being competitors in the markets. This shows that the configuration of the automotive industry is causing major change with the aim of more efficient competition.

Lately, in the new adoption of electric and hybrid car models, new industries may start to be considered competitors for traditional automakers. For instance, Tesla Motors, a well-known brand that manufactures electric cars has consolidated cooperation with Daimler (within the Mercedes-Benz brand) and Toyota. The expansion of what a competitor means would also reshape the scope of analysis for a coopetition strategy.

Not to mention that some initial cooperation plans ended in the absorption of one of the partners by the other, or in partial control. Nissan and Renault cooperated until the French carmaker bought almost half of Nissan. The Japanese group is still an autonomous player within the Nissan–Renault Alliance (Segrestin 2005, reviews this partnership). While still independent, they develop some cross investment in line with the interests of the other. They declare economies of scale to be the underlying reason for carrying out such an alliance. Nowadays, this agreement represents about one tenth of worldwide car sales. In 2010, they also joined forces with Daimler in order to enhance these sharing practices.

Hence, on the basis of this developing phenomenon, we not only need to research the outcomes but also the potentials. Managers and policymakers can be assisted by better analysis of coopetition potentials. This is the purpose of the following Section, where an economic approach to this subject is developed.

11.3 Methodology

11.3.1 Potential Coopetition

We introduce some notation and an analytical framework within which to study potential coopetition among plants in the automobile industry composed by I plants, indexed $q = 1, \ldots, I$. The output quantity and price vectors of a plant are given by $y = (y_1, \ldots, y_M) \in R^M_+$ and $p = (p_1, \ldots, p_M) \in R^M_{++}$, and its input quantity and price vectors by $x = (x_1, \ldots, x_N) \in R^N_+$ and $w = (w_1, \ldots, w_N) \in R^N_{++}$. Total assets of a plant are expressed by $A \in R_{++}$, which can differ from the input capital depending on its accounting definition. The profit is given by $\pi = R - C = p^T y - w^T x$, where "T" represents the transpose of the vector and, additionally, $w^T x = c^T y$ where $c = (c_1, \ldots, c_M) \in R^M_{++}$ defines the vector of unitary costs. The return on assets (ROA) of a plant is expressed by the ratio of profit to assets, π/A. The set of technologically feasible combinations of output vectors and input vectors is defined by the mathematical programming model known as Data Envelopment Analysis (DEA) introduced by Banker et al. (1984).

$$T = \left\{ (x, y) : y \leq \sum_{q=1}^{I} \lambda_q y_q, x \geq \sum_{q=1}^{I} \lambda_q x_q, \sum_{q=1}^{I} \lambda_q = 1, \lambda > 0 \right\}. \tag{11.1}$$

The representation of the technology in terms of its output set is $P(x) = \{y: (y, x) \in T\}$, which is bounded above by the output isoquant. Shephard (1970) introduced the output distance function, which provides a radial measure of the distance from an output vector to the output isoquant. This is defined as $D_O(x, y) = \min \{\mu: y/\mu \in P(x)\} \leq 1$. The output distance function is interpreted as a measure of the technical efficiency of a plant. There is efficiency when $D_O(x, y) = 1$. Otherwise, the plant is considered technically inefficient and its degree increases with lower values departing from one.

For simplicity, the exposition that follows is based on the potential coopetition between plants h and l. The methodology can easily be extended to a situation of coopetition between multiple plants.

Coopetition, in contrast with a merger, maintains the independence of the two plants introducing flexibility to the cooperation. A plant can easily switch the cooperation from one plant to another to seek the highest possible return on its investment. This is the economic incentive for coopetition, a behavior that is only possible if the plant maintains control of its own investment as well as the rest of its inputs. Hence, the aggregate assets and the aggregate inputs associated with coopetition between plants h and l are simply the sum of their quantities ($A_h + A_l$, $x_h + x_l$). We consider this to be feasible when: $(x_h + x_l, y_h + y_l) \in T$, being $y_h + y_l$ the aggregation of the output quantities of the two plants. The potential joint product as a result of coopetition is given by y_{h+l}, where $y_{h+l} = (y_h + y_l)/D_O(x_h + x_l, y_h + y_l)$. Therefore, the potential joint product is the maximum possible

given $x_h + x_l$ or the efficient one associated with the aggregate input quantities. The two firms translate the gains of efficiency from coopetition to a higher amount of output.[3] Thus, all possible complementarities from coopetition are captured when moving from independent to cooperative operations. The potential joint profit is $\pi_{h+l} = p^T_{h+l} y_{h+l} - w^{h+lT}_h x_h - w^{h+lT}_l x_l$ and $R_{h+l} = p^T_{h+l} y_{h+l}$ defines the potential joint revenue, where p_{h+l} is the new vector of prices associated with the potential joint product (y_{h+l}) and $w^{h+l}_k, k = h, l$ are the prices associated with each plant's quantity of inputs. A variation in the prices of inputs is not expected because the pressure of the plant on suppliers has not changed. The potential return on assets is defined as $ROA_{h+l} = \pi_{h+l}/(A_h + A_l)$.

The present study of potential coopetition between European automobile plants is based on the returns on assets, which is a well-known measure of financial performance.[4] This measure has the virtue of being independent from plant size. This property, which is not shared by other measures of financial performance, such as profit, revenue and cost, makes ROA particularly suitable for the study of potential coopetition between plants, which may be of disparate sizes.[5] We define a situation of potential coopetition as

Definition There is a potential economic incentive for coopetition between plants h and l, when $ROA_{h+l} > ROA_{k,}$ $k = h,l$ and $ROA_{h+l} \geq 0$.

The described situation is only possible when: $\pi_{h+l} > \pi_h + \pi_l$ and $\pi_{h+l} \geq 0$, the potential nonnegative joint profit is higher than the aggregation of profits from the individual plants. It is also interesting to note that potential coopetition implies that each participant receive a positive (but not necessarily equal) share of the gain, i.e. share of the potential joint revenue (R_{h+l}). In fact, if α_k, $k = h,\ l$ defines the proportion of the potential joint revenue that plant k receives (where $\sum_k \alpha_k = 1$), the potential coopetition involves: $\frac{\pi^{h+l}_{\alpha_k}}{A_k} > \frac{\pi_k}{A_k}, k = l, h$ where $\pi^{h+l}_{\alpha_k} = \alpha_k (p^T_{h+l} y_{h+l}) - w^{h+lT}_k x_k, \quad k = l, h$. In other words, the possibility of a sole player being able to appropriate all of the gains from coopetition is dismissed.

[3]The previous definition of potential joint product from coopetition scales each output with $D_O(x_h + x_l, y_h + y_l)$. As we have seen, coopetition may only affect some of the products. In this case, outputs should not be treated symmetrically and scale only some of them. The ones object of coopetition.

[4]See Chap. 8 of Grifell-Tatjé and Lovell (2015) for a comprehensive introduction to this measure of financial performance.

[5]Lozano (2013) takes a DEA-cost approach and seeks to minimize the cost of the planned joint-venture facility.

11.3.2 Decomposing ROA Change from Potential Coopetition

It is relevant to study the drivers of this potentially superior return on assets as a result of coopetition. We take a well-established approach in the business literature known as the duPont triangle (Johnson 1975), whereby the rate of return on assets is expressed as the product of two components, the profit margin and the assets rotation, i.e. ROA = π/R × R/A. The distinction between a situation of potential coopetition and non-coopetition can be expressed as

$$\frac{ROA_{h+l}}{ROA_k} = \frac{\pi_{h+l}/R_{h+l}}{\pi_k/R_k} \times \frac{R_{h+l}/(A_h + A_l)}{R_k/A_k}, \quad k = h, l. \tag{11.2}$$

The existence of potential coopetition has its origin in a better profit margin (higher profit by unit of revenue) and/or faster asset turnover. The first term on the right side of expression (11.2) takes a higher, equal or lower value than one in which the potential profit margin from coopetition is higher, equal to or lower than a situation of non-coopetition. There are two possible explanations: (i) divergence in prices and; (ii) different output–input relationship. Higher revenue per unit of assets is the effect that explains a faster asset turnover in the second term on the right side of (11.2). We pay attention to the profit margin component of the duPont triangle. We have

$$\frac{\pi_{h+l}/R_{h+l}}{\pi_k/R_k} = \frac{\pi_k^{h+l}/R_k^{h+l}}{\pi_k/R_k} \times \frac{\pi_{h+l}/R_{h+l}}{\pi_k^{h+l}/R_k^{h+l}}, \quad k = h, l, \tag{11.3}$$

where $R_k^{h+l} = p_k^T y_{h+l}$ expresses potential joint revenue and $\pi_k^{h+l} = p_k^T y_{h+l} - w_k^T x_h - w_k^T x_l$ potential joint profit with the prices of plant $k = l, h$. The two terms on the right side of expression (11.3) have a clear interpretation (Grifell-Tatjé and Lovell 2015, pp. 350–351). The first is a productivity effect and measures the potential contribution to the profit margin of changes in the level of productivity from a situation of non-coopetition to one of coopetition. The second is a price recovery effect and quantifies the potential impact of price variation on margin. Both expressions can take higher, equal or lower values than one showing productivity (price recovery) improvement, stagnation or decline. Additionally, Grifell-Tatjé and Lovell (2015) have shown how the productivity effect component in (11.3) can be decomposed; which is the approach that we take.

The potential joint profit with the prices of plant k can be re-expressed as $\pi_k^{h+l} = (p_k - c_k^{h+l})^T y_{h+l}$, where $c_k^{h+lT} y_{h+l} = w_k^T (x_h + x_l)$, $k = h, l$. It allows to write the potential profit margin from a situation of coopetition as

$$\frac{\pi_k^{h+l}}{R_k^{h+l}} = \left[\frac{p_k - c_k^{h+l}}{R_k^{h+l}}\right]^T y_{h+l}, \quad k = h, l$$
$$= \rho_k^{h+lT} y_{h+l}, \quad k = h, l, \tag{11.4}$$

and, in a similar way, the profit margin associated with the situation of non-coopetition can be expressed as $\pi_k/R_k = \rho_k^T y_k$ where $\rho_k = (p_{k1} - c_{k1}, \ldots, p_{kM} - c_{kM})/R_k$, $k = h, l$ defines an unitary margin expressed in prices of k. The productivity component on the right side of expression (11.3) can be rewritten using the previous results as

$$\frac{\pi_k^{h+l}/R_k^{h+l}}{\pi_k/R_k} = \frac{\rho_k^{h+lT} y_{h+l}}{\rho_k^T y_k}, \quad k = h, l, \tag{11.5}$$

and the direct application on (11.5) of the definition of potential joint production from coopetition: $y_{h+l} = (y_h + y_l)/D_O(x_h + x_l,\ y_h + y_l)$ enables us to re-express (11.5) as

$$\frac{\pi_k^{h+l}/R_k^{h+l}}{\pi_k/R_k} = \frac{\rho_k^{h+lT}(y_h + y_l)}{\rho_k^T y_k} \times \frac{1}{D_O(x_h + x_l, y_h + y_l)}, \quad k = h, l \tag{11.6}$$

Figure 11.1 depicts the decomposition of this expression (11.6). It represents the set of technologically feasible combinations of output and input quantities for the case of M = N = 1. It also shows the output and input quantities of plants h and l, which are located on the interior of the DEA technology. Hence, Fig. 11.1 illustrates a general situation in which an automobile plant can be inefficient, i.e. it is not on the frontier of the technology. It also depicts the aggregation of input and output quantities of the two plants: $x_h + x_l$, $y_h + y_l$, as well as the potential joint product from coopetition (y_{h+l}), which is located on the production frontier. The first term on the right side of expression (11.6) quantifies, in terms of potential profit margin change, the movement from (x_k, y_k) to ($x_h + x_l$, $y_h + y_l$) in Fig. 11.1. This can be considered the starting point, and is the result of a passive coopetition. It can take a value higher, equal or lower than one. The second term collects, in fact, the potential fruits from coopetition and measures, also in terms of profit margin change, the movement from ($x_h + x_l$, $y_h + y_l$) to ($x_h + x_l$, y_{h+l}) in Fig. 11.1.

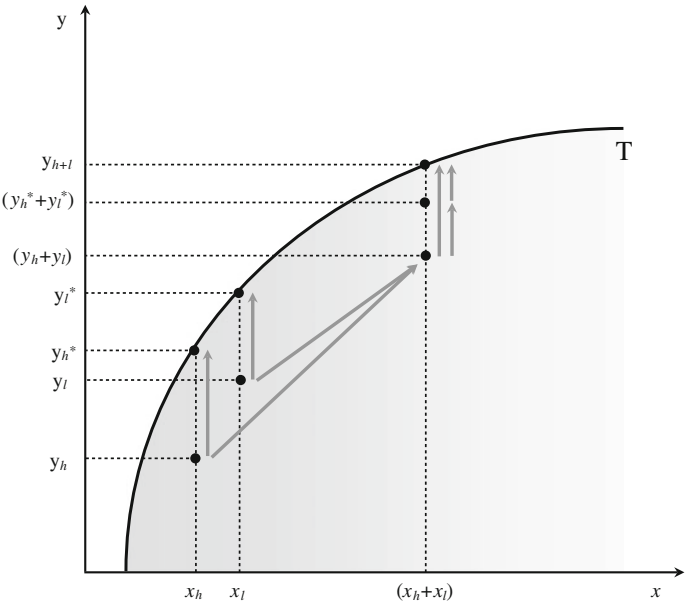


Fig. 11.1 Decomposition of coopetition effects

The decomposition of profit margin change in expression (11.6) can be linked with the previous work by Bogetoft and Wang (2005), who coined the second term on the right side of expression (11.6) *potential overall gains*. They consider it to comprise a portion of gain that could be achieved individually, before any sort of interaction between the units. That is to say, plants, prior to coopetition, could improve their operations in a way that enables them to achieve the best practices in the technology. They could reach their benchmarks before any sort of achievement from coopetition, i.e. $y_k^* = y_k / D_O(x_k, y_k), \quad k = h, l$. In terms of Fig. 11.1, this involves the movement from (x_k, y_k) to $\left(x_k, y_k^*\right), \quad k = h, l$. From this point of view, the term that Bogetoft and Wang (2005) called "potential overall gains" should first be adjusted in order to correctly evaluate the potential contribution of coopetition. The main idea is to evaluate only the improvements that cannot be reached individually as potential gains from coopetition, which implies the decomposition of the second term on the right side of expression (11.6) as follows:

$$\frac{\pi_k^{h+l}/R_k^{h+l}}{\pi_k/R_k} = \frac{\rho_k^{h+lT}(y_h + y_l)}{\rho_k^T y_k} \times \frac{D_O\left(x_h + x_l, y_h^* + y_l^*\right)}{D_O(x_h + x_l, y_h + y_l)} \times \frac{1}{D_O\left(x_h + x_l, y_h^* + y_l^*\right)}, \quad k = h, l, \quad (11.7)$$

where $y_{h+l} = (y_h^* + y_l^*)/D_O(x_h + x_l, y_h^* + y_l^*)$ and the second term on the right side of expression (11.7) quantifies the part of potential overall gains that can be reached individually, i.e. without any kind of coopetition. This is reflected in Fig. 11.1 by the movement from $(y_h + y_l)$ to $(y_h^* + y_l^*)$. At this point, the third term measures the contribution of all the potential achievements that are merely due to coopetition, as individual improvements prior to interaction are removed. This situation corresponds with the movement from $(y_h^* + y_l^*)$ to (y_{h+l}) in Fig. 11.1. This expression can be coined as *pure* coopetition effect. Note that this third term can take a value higher, equal to or lower than one. If the third term takes a value lower than one, it means that $(y_h^* + y_l^*) \notin P(x)$, the same as $D_O(x_h + x_l, y_h^* + y_l^*) > 1$. There are no gains associated with coopetition because plants can potentially reach a better level of profit margin alone, with self-adjustments. We refer to this movement as the technical efficiency effect.

As a brief summary, expression (11.7) proposes a decomposition of productivity difference based on three components, which is completed by the price effect in expression (11.3) and the asset turnover in expression (11.2). The product of these five effects gives a complete explanation of potential ROA gains between a situation of coopetition and non-coopetition.

11.4 The Data Set

It is worth noting that the purpose of this chapter is to study, from an economic perspective, the potential cooperation among independent automobile production plants from 2000 to 2012 inclusive. The automobile sector is one of the main contributors to the economy in the EU, as well as worldwide, and one of the largest providers of employment. Eurostat and the "Association des Constructeurs Européens d'Automobiles" (ACEA) reported that 2.2 million people were directly employed in the EU automobile sector in 2012. This figure rises to more than 3 million people when indirect employment is included. Since important policy measures were undertaken at the EU-level during the 2000–2012 period, the sample is limited to plants that are part of the EU-28. For this group of European countries, the regulatory environment is considered to be more similar and standardized. Not all countries were permanent EU-28 members since the year 2000. However, all of them had been official candidates since at least 1997. Croatia[6] is the only exception, whose candidacy was made official in 2004 and which became a member in 2013.

This study works with plant-level data and the sample is drawn from the European Automobile Manufacturing industry. The main source of information is the Amadeus database, which collects multidimensional accounting information from European automobile manufacturing companies. Specifically, the sample was

[6]Only one plant in the sample is located in Croatia.

extracted from the NACE[7] code 2910 titled "*manufacture of motor vehicles*" of the Amadeus database.[8] The long period of this study, 2000–2012, was characterized by major changes in the economic environment, which undoubtedly had some impact on the industry under analysis. The sample contains both private and public production plants although the latter are a small minority.

The Amadeus database provides financial information on individual production plants, our unit of analysis. Plants generate and provide their own accounting records, i.e. balance shit and income statement. In order to study relevant observations, plants whose average number of employees during the period was lower than one hundred were ignored. Furthermore, plants whose data was unreasonable or inconstant during the period of analysis were also dropped.[9] The transition from local general accepted accounting principles (GAAP) to international financial reporting standards (IFRS) was a matter of special attention in producing the dataset. This transition was slightly progressive from 2005 onwards.

The final dataset consists of 160 production plants belonging to 18 European countries and some of these production plants belong to the most important automobile production groups. The dataset takes an unbalanced panel-data configuration. There are plants without available information for one or more years. But also, aside availability and screening mentioned above, a high birth and death rate during the period helps to explain the unbalanced panel-data configuration. Offshoring processes carried out in the last decade surely explain a large part of the high birth and death rates observed.[10]

The amount of profits in a period of time is given by the accountancy records of the plant. These also account for the investment in assets. In accounting argot, these profits are referred to as "*earnings before taxes*" (EBT). This applied part follows a value added approach because information about the quantity of the intermediate materials is not available or insufficient. What is detailed in the accountancy records is the total cost of the period associated with intermediate materials. Hence, value added is defined by the total revenues minus the total cost of these intermediate materials. In this value added approach, two inputs are considered: labor and capital. This implies that revenues (R) are equal to the value added in the application. We describe and name the relevant variables for inputs (labor and capital) and output (value added) as follows:

[7]"Statistical Classification of Economic Activities in the European Community", subject to legislation at the EU level, which imposes the use of the classification uniformly across all Member States.

[8]Data download/collection took place twice between 2011 and 2013. Thus, the criteria for unit selection was their main activity (NACE: 2910) available at the time of download.

[9]Some plants were removed from the sample because abnormal trends for some relevant variables were found e.g. number of employees, amount of assets, compensation per employee, price of capital, among others.

[10]The traditional definition of offshoring includes both the practice of a unit hiring external functions from a third party—outsourcing—and the case of moving to a different location, which explains both the birth and death of plants in the sample.

(i) *Labor quantity* (x_1). The quantity of labor is defined as the average number of employees of the plant during the year. This is computed as the average of the total reported number of employees at the beginning of the accounting period and at the end.

(ii) *Labor price* (w_1). This is defined by the ratio between the total labor compensation of the plant and labor quantity. Consequently, the product of labor quantity and its price is equal to the total labor cost for the plant during the period.

(iii) *Capital quantity* (x_2). The starting point is the value of the net tangible fixed assets in the plant's accounting records in the year 2000 (x_2^{2000}). To construct the capital stock of the following year (2001), the annual assets depreciation of the year is first subtracted from the capital stock existing at the beginning of the period. This can be expressed as: $x_2^{2000}(1-\delta^{2001})$, where δ^{2001} expresses the depreciation rate for the period. Second, the investment made by the company during the year 2001 (I^{2001}) is identified. Third, this investment is valued at constant 2000 prices by applying the consumer price index of that plant's country as a deflator, i.e. $I^{2001}/(1+d_{2000}^{2001})$ where d_{2000}^{2001} represents the consumer price index of period 2001. Fourth, the stock of capital of period for 2001 is defined by the sum of this deflated investment plus the previously calculated adjusted assets of 2000 $(x_2^{2000}(1-\delta^{2001})+I^{2001}/(1+d_{2000}^{2001}))$. The capital stock for the following year 2002 is calculated in exactly the same way and so on for the remaining years. In summary, the capital stock for the year $t+1$ is calculated as $x_2^{t+1} = x_2^{t}(1-\delta^{t+1})+I^{t+1}/(1+d_{2000}^{t+1})$, $t = 2000, \ldots, 2011$ where d_{2000}^{t+1} is the cumulative deflator from 2000 to year t + 1.

(iv) *Capital price* (w_2). This is calculated as the ratio between total capital costs of the plant (interest paid plus depreciation) and the capital stock for the period. Therefore, the product of capital quantity and its price is equal to the plant's total capital costs for the period.

(v) *Product quantity* (y). This is expressed as the plant's constant value added. Its value added for the period is deflated by a cumulative manufacturing producer price index for the domestic market (base 2000) where the plant is located. This is expressed at constant 2000 prices. The *output price* (p) is defined by the ratio between the value added for the period and the product quantity (y). Thus, the output price is the cumulative manufacturing producer price index for the domestic market (base 2000) of the country where the plant is located.

(vi) *Total Assets* (A). The amount of total assets is taken from the plant's accountancy statements.

Table 11.1 shows the mean values per each variable for the 160 plants in the final dataset. Moreover, it presents two different periods in order to observe changes or some sort of trend in the sample configuration. The start of the global financial crisis (2007–2008) is taken to segment the data into two subsamples: 2000–2007

Table 11.1 Summary statistics per variable. Mean values

Period	x_1	w_1	x_2	w_2	y	p	Profit (current €)	A	ROA (%)	ROA median (%)
	Labor quantity (#)	Labor price (current €)	Capital quantity (€2000)	Capital price (current €)	Product quantity (€2000)	Product price index		Total assets (€2000)		
2000/12	1961	40,013	155,935,690	0.3631	118,201,157	114.62	13,261,665	478,149,989	3.70	3.17
2000/07	2078	39,039	170,120,065	0.3476	133,597,155	106.92	16,720,242	476,261,088	4.58	3.32
2008/12	1819	41,207	139,978,268	0.3809	96,750,919	126.95	8,549,003	480,724,670	2.50	2.85

and 2008–2012. An average plant size of nearly two thousand employees is found. However, the situation changes notably per period: there is a reduction of almost three hundred workers, on average, between the first and second periods. More in depth analysis has shown that the average decline rate per year was over 2.71%, with a more intense drop in the first half. Both capital quantity and product quantity present a somewhat similar pattern. The trend of reduction overlaps with what has been expressed regarding labor quantity, and the decline rates are rather similar for the period (somewhat stronger for product quantity). Total assets, however, present slight growth. Aside from this latter point, it can be argued that there was a tendency to downsize in this industry between 2000 and 2012. As for prices, labor price increased slightly. This may reveal a convergent trend in Europe, as this increase may be motivated by a faster rise in wages in some peripheral countries. Capital price increased slowly with an inverted U-shape throughout the whole period. Profits also decreased, collapsing to half by the second period. It is worth mentioning that this is mainly due to a global loss in year 2009, when average profit was negative. From 2010 on, plants seem to make an effort to control and adjust their costs, despite the ongoing declines in the markets. Regarding product price, as a deflator is being used, this only shows the accumulated value of the producer price index as stated above.

Finally, return on assets present an average value of 3.7% in our sample, which mimics the typically stated for this industry, between 3 and 5%. However, it is also true that the mean values conceal an inverted U-shape of this magnitude with a clear drop in the period of crisis and only a shy recovery in recent years. Table 11.1 also shows median ROA values. Some upcoming tables are shown in median values, which are presented for better understanding. Median values for ROA are somewhat different from mean results. However, they depict a very similar trend, especially in the crisis period.

11.5 Results

11.5.1 Two-Plants Interactions

In this section, potential coopetition between European automobile production plants is evaluated. Our data sample allows for the construction of 45,332 valid interactions throughout the 13-year period,[11] during which 160 plants participated in at least one interaction. However, the observations for the analysis were selected in accordance with a set of criteria. The first criterion involved using only those combinations laying inside the technology in the original projection ($y_h + y_l \in P(x)$). Second, according to the definition of potential economic incentive for plants

[11]In the application, as some of the variables were built as mean values between the beginning and the end of the period, the study eventually worked with 12 periods instead of 13.

to take part in coopetition, cases were only considered if the coopetition offered an improvement on ROA for both of the plants involved ($ROA_{h+l} > ROA_k$, $k = h, l$). That is, coopetition offered an economic improvement to both actors. Third, and as made implicit in the meaning of coopetition, observations were eliminated when the cooperating plants belonged to the same producer or group, as the study only focused on competing plants.

So, the corpus was narrowed down to 34,080 cases of potential coopetition. All of these offered potential economic gains from a coopetition strategy. It should be noted that, out of the total number of potential cases of coopetition, there were 15,195 cooperations (more than 44% out of 34,080) in which at least one of the plants became viable: it started with a negative ROA and the potential ROA_{h+l} was potentially positive.

However, yet another criterion needed to be met, which is related to the adjustment that plants could make individually, before any interaction. Following expression (11.7) in the methodological section, additional cases were removed when all possible gains could be achieved by individual efforts and the effect of pure cooperation did not contribute, i.e. $\left(y_h^* + y_l^*\right) \notin P(x)$. This last step led to a final sample of 12,241 cases[12] of potential coopetition (roughly, more than a fourth of the total initial possible interactions). The analysis that follows is based on these cases.

11.5.2 Exploring Potential Coopetition

In order to gain insight into the configuration of these interactions, the initial financial performance of the plants was first analyzed. Production plants before cooperating could perform with a positive or a negative ROA. So, there were three possibilities of coopetition: cases where both plants had a positive ROA before cooperating, cases where both plants had a negative one and cases where one of them had a positive one whereas the other was negative. Table 11.2 shows the results for these three possibilities. In Table 11.2, the 2000–2012 period has been divided into two sub-periods: 2000–2007 and 2008–2012, which correspond to before and during the economic crisis, respectively. The results are also shown for these two periods of time, including both the percentage and the number of cases. Percentages are shown per row.

[12]These 14,933 cases are almost equally distributed between two possible periods of time: before and during the economic crisis. This is, from 2000 to 2007 there are 7736 possible cases of coopetition and from 2008 to 2012 there are 7197 (51.8 and 48.2%, respectively).

Table 11.2 Distribution of cases according to initial ROA status of plants

Period/status	Both positive	One positive, one negative	Both negative	Total
2000/12	57.60% (7,051)	36.20% (4,431)	6.20% (759)	100% (12,241)
2000/07	63.54% (4,544)	32.14% (2,298)	4.32% (309)	100% (7,151)
2008/12	49.25% (2,507)	41.91% (2,133)	8.84% (450)	100% (5,090)

As can be observed, there is a tendency of change between the two periods. In both periods, cooperations between both plants presenting an initially positive ROA dominate (63% and almost 50%, respectively per period). Potential cases in which plants could enter with a different status are also relevant: almost one third in the first period and roughly 42% in the second half.

This change is a consequence of the economic environment in the second half. Cases of potential coopetition in which both plants start with negative ROA represent a minor portion, but it is also true that it more than doubles in the second half, approaching to a tenth of the cases. As previously pointed out, following the definition in the methodological section, the final outcome of the potential coopetition must be a non-negative ROA.

Table 11.3 Distribution of coopetition cases according to size of plants. Number of cases in brackets

Size	Big	Medium–big	Medium–small	Small
Big	0.20% (24)	4.77% (584)	4.20% (514)	14.93% (1,828)
Medium–big		4.57% (560)	8.37% (1,025)	18.50% (2,264)
Medium–small			7.52% (921)	24.88% (3,046)
Small				12.05% (1,475)
Total	0.20% (24)	9.35% (1,144)	20.10% (2,460)	70.36% (8,613)

Another area of interest focuses on the size of plants involved in coopetition, as shown in Table 11.3. Percentage values are calculated out of the total 12,241 potential coopetition cases, so that the sum of percentages in the table corresponding to the ten possible combinations of size amounts to 100%. The classification of sizes was carried out in accordance with quartile values of the number of

employees per year.[13] Table 11.3 shows that potential coopetition occurs more between different-sized plants. By adding up the diagonal in the table, where cooperating plants are categorized with the same size, the number is less than a quarter of the total number of cases. If the last two columns are observed, it can be seen how the percentages amount up to more than 90%, showing that a large proportion of the potential cooperations corresponds to small and medium–small plants. More interestingly, this fraction is already 70% if only the last column referring to small plants is considered. In a later table, these results will be considered in light of the study.

11.5.3 ROA Gains and Drivers

In this section, potential change in ROA is analyzed as well as the main drivers for this change. Table 11.4 shows both the median value for ROA gains and its decomposition. It presents median values, instead of mean ones, due to the frequent generation of extreme results. These are motivated by the fact that some plants have a very low starting ROA, so that a moderate potential ROA would produce an extreme ROA change. In this situation the median is more informative. Recall that Eqs. (11.2), (11.3) and (11.7) do not hold in Tables 11.4 and 11.5 because of this median approach. Columns three and four in Table 11.4 show a ROA decomposition based on Eq. (11.2), in which a faster asset turnover or/and a better profit margin explains a higher ROA from coopetition. Computed as (ROA_{h+l}/ROA_k) in Eq. (11.2), the ROA would potentially improve nearly five times and profit margin and assets turnover seem to contribute in equal rather terms. However, there is some difference when analyzing per period. A considerable reduction in the potential ROA gains of one and a half points can be observed. This drop has its origin in a reduction in the profit margin that is not compensated by the assets turnover for the 2008/12 period. A deeper observation of the results shows a quite constant assets turnover change over the years, whereas the profit margin change is less stable and presents lower values in the second half. Again, the results are clueing that the plant's results have been affected by the so-called economic crisis in Europe and profit margins have fallen, even when they are described in potential terms.

Columns five and six in Table 11.4 are based on Eq. (11.3), showing whether changes in the profit margin originate from changes in prices or productivity. The result in column five indicates that prices are practically neutral, it thus being the productivity effect that actually drives the change. Therefore, all the potential profit margin changes moving from individual to joint production come from productivity

[13]Number of employees ranges from 100 to 14,890. Quartiles for size distribution, calculated per year, vary slightly from year to year. Thus, mean intervals for the distribution are (100; 245), (245; 627), (627; 2,844) and over 2,844 for 'small', 'medium-small', 'medium-big' and 'big', respectively.

Table 11.4 Decomposition of potential ROA gains. Median values

Period	ROA Gains	=	ROA Gains		Profit Margin Change		Productivity Effect		
			Assets Rotation Change	x Profit Margin Change	Price Recovery Effect	x Productivity Effect	*Passive* Coopetition Effect	x Technical Efficiency Change	x *Pure* Coopetition Effect
[1]	[2]		[3]	[4]	[5]	[6]	[7]	[8]	[9]
2000/12	4.796		2.215	2.340	1.005	2.180	1.062	2.053	1.051
2000/07	5.518		2.257	2.489	1.005	2.320	1.040	2.100	1.047
2008/12	3.968		2.151	2.146	1.005	2.018	1.087	1.999	1.057

gains. Hence, it can be argued that the potential reduction in profit margin change between periods is caused by lower potential productivity gains.

The productivity effect is further decomposed into three other drivers expressed in Eq. (11.7). Columns seven to nine in Table 11.4 show these results. The main finding to be highlighted is that the productivity effect, and therefore the profit margin change, is highly determined by the so-called technical efficiency effect. This is, higher profits per unit of revenue are likely to be achievable with individual efforts, if plants are able to imitate better practices in the industry. This effect was expected, and remarkable values of technical inefficiency for some of the plants in the original sample were found. Consequently, such a result is yet not surprising as some literature has already found the automobile industry to be traditionally shaped by some 'mediocre survivors' (Holweg and Oliver 2016).

Passive coopetition shows a non-trivial contribution to ROA gains. Column nine depicts the effect on productivity of pure coopetition, in terms of profit margin changes. It shows a relevant positive impact too, vaguely superior in the second half. Being less important than the technical efficiency change, the natural question that arises is whether plants can achieve an impressive improvement in the level of efficiency (column eight) by themselves without coopetition. If this is not the case, the technical efficiency change must fairly be considered one of the outcomes of coopetition. They need a cooperation agreement as an incentive for their own reorganization although coopetition produces a reduced additional joint product. In this case, the technical efficiency change might be considered together with the passive coopetition effect.

The idea signaled in Table 11.3 is further developed in Table 11.5. Following the same definition of size as before, ROA gains, as well as their drivers, are shown for different categories of plant size.

Table offers a result that we may analyze in two steps. Initially, we see that the smaller the plant the lower the potential ROA gains. While bigger plants would potentially increase their ROA 5.4 times, smaller ones would increase it 4.5 times. For bigger plants, ROA gains are accelerated by higher profit margin changes whereas smaller ones achieve potentially faster assets rotation. In all cases, this is fully driven by productivity gains and, as in the general case, the main source is the technical efficiency change. As for passive coopetition, it is found it to offer rather divergent results. Biggest group of plants get the most out of this stage, but smallest

Table 11.5 Decomposition of potential ROA gains, per plant size. Median values

				ROA Gains		Profit Margin Change		Productivity Effect		
Period	Plant Size	# Obs	ROA Gains =	Assets Rotation Change x	Profit Margin Change	Price Recovery Effect x	Productivity Effect	*Passive* Coopetition Effect x	Technical Efficiency Change x	*Pure* Coopetition Effect
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
2000/12	B	2,974	5.445	2.069	2.611	1.001	2.578	1.430	2.030	1.010
	MB	4,993	5.062	2.191	2.334	1.001	2.239	1.040	2.110	1.036
	MS	6,427	4.752	2.292	2.200	1.037	2.015	1.030	2.055	1.091
	S	10,088	4.534	2.254	2.364	0.997	2.257	0.952	2.032	1.092
2000/07	B	2,163	5.646	2.039	2.644	1.002	2.615	1.360	1.975	1.012
	MB	2,886	5.476	2.284	2.417	1.000	2.333	0.995	2.192	1.036
	MS	3,721	5.254	2.313	2.262	1.039	2.036	1.076	2.077	1.088
	S	5,532	5.601	2.384	2.537	0.992	2.444	0.889	2.120	1.096
2008/12	B	811	4.429	2.179	2.407	0.999	2.163	1.558	2.140	1.005
	MB	2,107	4.647	2.045	2.258	1.001	2.111	1.098	2.002	1.037
	MS	2,706	3.715	2.261	2.017	1.034	1.809	0.970	2.027	1.094
	S	4,556	3.670	2.103	2.125	1.007	1.963	1.028	1.953	1.087

B: big; MB: medium–big; MS: medium–small; S: small

ones find an unfavorable result, as a value lower than one would potentially cause ROA losses. This is justified by the fact that bigger plants present lower values of starting ROA than smaller plants. However, if we have a look at pure coopetition effect, we conversely find that the smaller the plant the higher the effect on ROA gains. Both results together point out that smaller plants must play an active role in coopetition to eventually obtain some gains. A passive interaction is detrimental for them.

Another peculiarity is found concerning periods. All categories worsen from 2008 onwards. However, the smaller the plant is, the higher the reduction in the potential ROA gains in the second half. For the smaller ones, this due to lower productivity gains, which makes the profit margin changes lower as well. The pure coopetition effect remains quite stable between the periods, keeping the same raking as in the general case.[14]

11.6 Conclusion

There is little literature covering the concept of coopetition, or cooperation among competitors, from an economic perspective. This chapter contributes to the field through the introduction of a non-parametric method to explore the potential economic gains from coopetition. Coopetition, rather than merging, offers many

[14]We have also carried out the same analysis per each of the ten types of interactions according to plant size (big to big, big to medium-big, etc.). Results emphasize the effect of size, as in Table 11.5. Passive coopetition only pays off for bigger plants whenever they interact with smaller partners. And pure coopetition effect is higher the smaller the plants taking part of the agreement, being small-small the best possible scenario.

advantages, flexibility being an important one of those. In fact, the plant maintains not only control of its own investment, but also the rest of productive factors. Furthermore, while merging would imply a permanent engagement between the plants, coopetition only happens when incentives pay off. In this context, the parties can terminate the cooperation if the conditions of the initial agreement are not upheld.

The chapter proposes a definition of potential economic incentives for coopetition between independent plants based on the rate of return on assets, a well-known measure of financial performance. The methodological approach has its roots in the previous work by Bogetoft and Wang (2005) and by Grifell-Tatjé and Lovell (2015) and enables comparison between situations of non-coopetition and coopetition. The methodology was applied to the study of potential competition within the European automotive industry. The main results are based on the analysis of a generated sample of over forty-five thousand cases of potential cooperation. Out of that sample, roughly twelve thousand of the cases (about 27%) showed potential ROA gains from coopetition.

The main findings reveal that faster asset turnover and better productivity explain higher potential ROA from coopetition. It makes a clear contribution to productivity gains, but the most important driver is technical efficiency. In theory, the plant can reach a higher level of efficiency by itself, without any kind of cooperation. However, the question is whether it needs a cooperation agreement as an incentive for its own reorganization. If that is the case, technical efficiency change should be considered to be an outcome of coopetition.

The results also show that medium–small and small plants have the strongest economic incentive for coopetition. However these groups of plants must play an active role in coopetition to get the potential gains. Passive coopetition would only be fruitful for bigger plants (they present lower ROA values) whereas smaller plants find it disadvantageous. This result seems natural, but empirical literature supporting this claim has not been found, which may be due to the legal framework in which coopetition is placed.

Results in the two periods defined as before and after the crisis are also significantly different. The period of financial distress, 2008–2012, presents lower ROA values and lower potential gains. That effect is more pervasive for smaller actors. It is also true that in the last years of this period appears an overall path to recovery.

Platform-sharing being a suitable method of coopetition in this industry, it has often been configured as a joint-venture between competing producers. In most cases, this arrangement is treated by law as a merger. When coopetition has a European dimension or is a full-function joint venture,[15] the EU regulation on

[15]Joint-ventures are regulated both by the EC Merger Regulation and Article 101 of The Treaty on European Union and the Treaty on the Functioning of the European Union. Joint ventures are virtually treated as merger-like operations. This link provides a summary of the assessment and treatment of joint-ventures under European Regulation: http://uk.practicallaw.com/1-107-3702# (last accessed February 2016).

mergers is applied. In other cases, special standards (Article 101), as well as EU or national competition authorities, must approve this type of agreements. Regardless, platform-sharing between plants must overcome many legal restrictions before finally being approved to operate.

Our results suggest that a specific regulation on coopetition needs to be issued at the EU level. Coopetition cannot be treated as a merger and it should be promoted instead of penalized. Policy makers should better understand the virtues of coopetition, removing the worry of hidden collusion. Actually, the new regulation should offer clear incentives for coopetition rather than preventing it, especially to medium and small plants. The potential gains from coopetition found are a good reason for this regulatory re-design, which can be achieved by issuing a specific legal framework.

Some limitations of this study make research extensions relevant in the applied side. For instance, as we discuss with regard to potentiality, many costs associated to the development of the coopetition strategy might reduce the gains to be captured. Further applications should consider some type of structure-, distance-, bargaining- or opportunity-related costs. The conclusions are based on the assumption that plants share gains, but this may not always be the case. The distribution may favor stronger plants, so smaller ones may not actually find such a favorable scenario in reality. Natural, potential extensions for future research on coopetition could also involve analyses about the effect on the consumers, market pricing, product range, product quality or overall surplus.

Acknowledgements The authors wish to acknowledge the financial support from the Spanish Ministry of Science and Innovation (Ref.: ECO213-46954-C3-2-R) and the FPI Scholarship Subprogram (Ref.: BES-2011-050619).

References

Alves J, Meneses R (2013) Partner selection in Co-opetition: a three step model. J Res Mark Entrepreneurship 17(1):23–35

Bagdadioglu N, Price CW, Weyman-Jones T (2007) Measuring potential gains from mergers among electricity distribution companies in Turkey using a non-parametric model. Energy J 28(2):83–110

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30(9):1078–1092

Bengtsson M, Kock S (2000) Coopetition' in business networks—to cooperate and compete simultaneously. Ind Mark Manage 29(5):411–426

Bengtsson M, Kock S (2014) Coopetition—Quo Vadis? Past accomplishments and future challenges. Ind Mark Manage 43(2):180–188

Biesebroeck JV (2007) Complementarities in automobile production. J Appl Econ 22(7): 1315–1345

Blum A (2009) Coopetition in the automotive industry. In: An investigation into the underlying causes, potential gains and losses. Ed. VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG

Bogetoft P, Wang D (2005) Estimating the potential gains from mergers. J Prod Anal 23(2): 145–171

Bouncken RB, Gast J, Kraus S, Bogers M (2015) Coopetition: a systematic review, synthesis, and future research directions. RMS 9(3):577–601

Bouncken R, Fredrich V (2012) Coopetition: performance implications and management antecedents. Int J Innov Manage 16(5):1–28

Chin KS, Chan BL, Lam PK (2008) Identifying and prioritizing critical success factors for coopetition strategy. Ind Manage Data Syst 4(108):437–454

Ehrenmann F, Reiss M (2011) Co-opetition as a facilitator of manufacturing competitiveness: opportunities and threats. In: ElMaraghy H (ed) Enabling manufacturing competitiveness and economic sustainability

Freyssenet M, Lung Y (2007) Car firms' strategies and practices in Europe. In: Faust M, Voskamp U, Wittke V (eds) European industrial restructuring in a global economy: fragmentation and relocation of value chains, vol 2004. SOFI Beritche, Göttingen, pp 85–103

Gnyawali DR, Park BJ (2011) Co-opetition between giants: collaboration with competitors for technological innovation. Res Policy 40(5):650–663

Grifell-Tatjé E, Lovell CAK (2015) Productivity accounting. The economics of business performance. Cambridge University Press, New York

Halkos GE, Tzeremes NG (2013) Estimating the degree of operating efficiency gains from a potential bank merger and acquisition: a DEA bootstrapped approach. J Bank Finance 37(5):1658–1668

Holweg M, Oliver N (2016) Crisis, resilience and survival. Lessons from the global auto industry. Cambridge University Press, Cambridge

Johnson HT (1975) Management accounting in an early integrated industrial: E. I. duPont de Nemours Powder Company, 1903–1912. Bus Hist Rev 49(2, Summer):184–204

Kristensen T, Bogetoft P, Pedersen KM (2010) Potential gains from hospital mergers in Denmark. Health Care Manage Sci 13(4):334–345

Li Y, Liu Y, Liu H (2011) Co-opetition, distributor's entrepreneurial orientation and manufacturer's knowledge acquisition: evidence from China. J Oper Manage 29(1–2):128–142

Lozano S (2013) Using DEA to find the best partner for a horizontal cooperation. Comput Ind Eng 66:286–292

Luo Y (2007) A coopetition perspective of global competition. J World Bus 42(2):129–144

Madhok A (1996) Know-how-, experience- and competition-related considerations in foreign market entry: an exploratory investigation. Int Bus Rev 5(4):339–366

Muffatto M (1999) Platform strategies in international new product development. Int J Oper Prod Manage 19(5–6):449–460

Oliver AL (2004) On the duality of competition and collaboration: network-based knowledge relations in the biotechnology industry. Scand J Manage 20(1–2):151–171

Pasche M, Sköld M (2012) Potential drawbacks of component commonality in product platform development. Int J Automot Technol Manage 12(1):92–108

Peng TJA, Bourne M (2009) The coexistence of competition and cooperation between networks: implications from two Taiwanese healthcare networks. Br J Manage 20(3):377–400

Ritala P, Sainio LM (2014) Coopetition for radical innovation: technology, market and business-model perspectives. Technol Anal Strateg Manage 26(2):155–169

Ritala P, Hurmelinna-Laukkanen P (2009) What's in it for me? Creating and appropriating value in innovation-related coopetition. Technovation 29(12):819–828

Rusko R (2011) Exploring the concept of coopetition: a typology for the strategic moves of the finnish forest industry. Ind Mark Manage 40(2):311–320

Segrestin B (2005) Partnering to explore: the Renault–Nissan alliance as a forerunner of new cooperative patterns. Res Policy 34(5):657–672

Shephard RW (1970) Theory of cost and production functions. Princeton University Press, Princeton

Walley K (2007) Coopetition: an introduction to the subject and an agenda for research. Int Stud Manage Organ 37(2):11–31

Wilhelm MM, Kohlbacher F (2011) Co-opetition and knowledge co-creation in Japanese supplier–networks: the case of Toyota. Asian Bus Manage 10(1):66–86

Wolff MF (2009) Automakers hope 'coopetition' will map route to future sales. Res Technol Manage 52(2):2–4

Zhang H, Shu C, Jiang X, Malter AJ (2010) Managing knowledge for innovation: the role of cooperation, competition and alliance nationality. J Int Mark 18(4):74–94

Zschille M (2015) Consolidating the water industry: an analysis of the potential gains from horizontal integration in a conditional efficiency framework. J Prod Anal 44(1):97–114

Chapter 12
Measuring Eco-efficiency Using the Stochastic Frontier Analysis Approach

Luis Orea and Alan Wall

Abstract The concept of eco-efficiency has been receiving increasing attention in recent years in the literature on the environmental impact of economic activity. Eco-efficiency compares economic results derived from the production of goods and services with aggregate measures of the environmental impacts (or 'pressures') generated by the production process. The literature to date has exclusively used the Data Envelopment Analysis (DEA) approach to construct this index of environmental pressures, and determinants of eco-efficiency have typically been incorporated by carrying out bootstrapped truncated regressions in a second stage. We advocate the use of a Stochastic Frontier Analysis (SFA) approach to measuring eco-efficiency. In addition to dealing with measurement errors in the data, the stochastic frontier model we propose allows determinants of eco-efficiency to be incorporated in a one stage. Another advantage of our model is that it permits an analysis of the potential substitutability between environmental pressures. We provide an empirical application of our model to data on a sample of Spanish dairy farms which was used in a previous study of the determinants eco-efficiency that employed DEA-based truncated regression techniques and that serves as a useful benchmark for comparison.

Keywords Eco-efficiency · Stochastic frontier analysis · Dairy farms

JEL codes C18 · D24 · Q12 · Q51

12.1 Introduction

Concerns about the sustainability of economic activity has led to an increasing interest in the concept of eco-efficiency and the literature on this topic has been growing in recent years (Oude Lansink and Wall 2014). The term eco-efficiency was originally coined by the World Business Council for Sustainable Development

L. Orea (✉) · A. Wall
Oviedo Efficiency Group, University of Oviedo, Oviedo, Spain
e-mail: lorea@uniovi.es

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_12

in their 1993 report (Schmidheiney 1993) and is based on the concept of creating more goods and services using fewer resources. In turn, the OECD defines eco-efficiency as "the efficiency with which ecological resources are used to meet human needs" (OECD 1998). Clearly, the concept of eco-efficiency takes into account both the environmental and economic objectives of firms.

When evaluating firm performance in the presence of adverse environmental impacts, production frontier models are a popular tool (Tyteca 1996; Lauwers 2009; Picazo-Tadeo et al. 2011; Pérez-Urdiales et al. 2015). The measurement of eco-efficiency in a frontier context, which Lauwers (2009) refers to as the 'frontier operationalisation' of eco-efficiency, involves comparing economic results derived from the production of goods and services with aggregate measures of the environmental impacts or 'pressures' generated by the production process. To date, only the non-parametric Data Envelopment Analysis (DEA) method has been used in the literature. While DEA has many advantages, it has the drawback that it can be extremely sensitive to outliers and measurement errors in the data.

In the present work we propose a Stochastic Frontier Analysis (SFA) approach to measuring eco-efficiency, which has the advantage that it well-suited to dealing with measurement errors in the data. Using a stochastic frontier model to measure eco-efficiency involves the estimation of only a few parameters, so the model can be implemented even when the number of observations is relatively small. Moreover, the SFA approach permits an analysis of the potential substitutability between environmental pressures and can incorporate determinants of eco-efficiency in a one-stage procedure.

We illustrate our simple proposal with an empirical application using a sample of 50 dairy farmers from the Spanish region of Asturias. This data set includes information from a questionnaire specifically carried out to permit the accurate measurement of eco-efficiency and provides information on farmers' socioeconomic characteristics and attitudes towards the environment, and has been used by Pérez-Urdiales et al. (2015) to measure eco-efficiency and identify its determinants using the DEA-based bootstrapped truncated regression techniques of Simar and Wilson (2007). The results from that paper therefore provide a useful point of comparison for the results from our proposed stochastic frontier model.

The paper proceeds as follows. In Sect. 12.2 we discuss the concept of eco-efficiency and the DEA approach often used to estimate eco-efficiency scores. Section 12.3 introduces our stochastic frontier model, which can be viewed as a counterpart of the DEA eco-efficiency model. Section 12.4 describes the data we use. The results are presented and discussed in Sect. 12.5, and Sect. 12.6 concludes.

12.2 Background

To measure eco-efficiency using frontiers, Kuosmanen and Kortelainen (2005) defined eco-efficiency as a ratio between economic value added and environmental damage and proposed a pressure-generating or pollution-generating technology set

$T = \{(\pi, p) \in R^{(1+K)} | \, \pi \text{ can be generated by } p\}$. This technology set describes all the feasible combinations of economic value added, π, and environmental pressures, p. Environmental damage, $D(p)$, is measured by aggregating the K environmental pressures $(p_1, \ldots, p_K)$ associated with the production activity.

Figure 12.1 provides an illustration for the simple case of two environmental pressures, p_1 and p_2. The set of eco-efficient combinations is represented by the eco-efficient frontier, which represents the minimum combinations of the two environmental pressures which can be used to produce an economic value added of π_0. Combinations of pressures below the frontier are unfeasible whereas combinations above it are eco-inefficient. For example, the combination of pressures represented by point A is clearly eco-inefficient as the environmental pressures could be reduced equiproportionally to point E on the frontier without reducing value added.

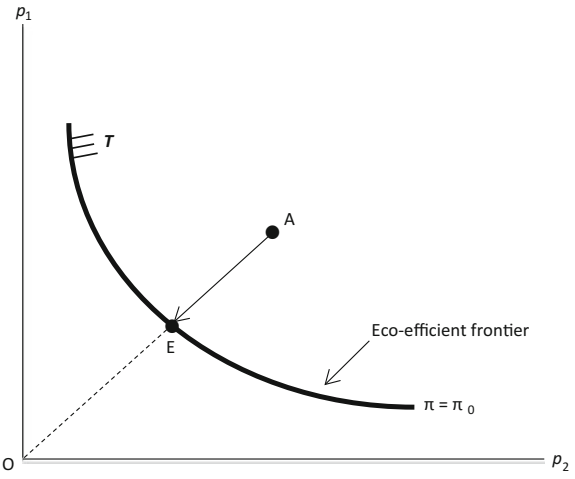


Fig. 12.1 Eco-efficiency

Eco-inefficiency can be measured using the radial distance from a point A to the efficient frontier. The eco-efficiency score is given by the ratio OE/OA which takes the value 1 for eco-efficient combinations of pressures and economic value added and values less than 1 for inefficient combinations such as A. This is the approach we will consider, although it should be pointed out that alternative measures of eco-efficiency could be devised if we depart from radial (equiproportional) reductions in pressures. For example, instead of measuring the extent to which pressures can be reduced while maintaining value added, we could measure the extent to which the firm, given its present combination of pressures, could increase its value added. Thus, if the firm was using the combination of pressure represented by A efficiently, it would be operating on a new eco-efficient frontier passing through that point, and could achieve a higher value added corresponding to this new frontier. Other alternatives exist where the possibility of simultaneously reducing pressures and increasing economic value added can be explored. Picazo-Tadeo et al. (2012), for example, propose using a directional distance function approach which allows

for this possibility, as well as that of reducing subsets of pressures in order to reach the eco-efficient frontier.

These different ways of approaching the eco-efficient frontier will all lead to valid measures of eco-inefficient behaviour but we will follow the existing literature by focusing on the capacity of firms to reduce environmental pressures equiproportionally while maintaining value added. It should be underlined that our eco-efficiency scores are defined directly in terms of environmental pressures and not in terms of reductions of input quatities which can be transformed into an associated reduction in in overall environmental damage. This latter approach was followed by Coelli et al. (2007) and permitted them to disaggregate environmental inefficiency into technical and allocative components using "iso-pressure" lines.

Individual eco-efficiency scores for producer i can be found using the following expression:

$$EEF_i = \frac{\text{Economic value added}}{\text{Environmental pressure}} = \frac{\pi_i}{D_i(p)} \tag{12.1}$$

where $D_i(p)$ is a function that aggregates the environmental pressures into a single environmental pressure indicator. This can be done by taking a linear weighted average of the individual environmental pressures:

$$D_i(p) = w_1 p_{1i} + w_2 p_{2i} + \cdots + w_K p_{Ki} \tag{12.2}$$

where w_k is the weight assigned to environmental pressure p_k. Kuosmanen and Kortelainen (2005) and Picazo-Tadeo et al. (2012), among others, use DEA as a non-subjective weighting method. The DEA eco-efficiency score of firm i can be computed from the following programming problem

$$\max_{w_{ki}} EEF_i = \frac{\pi_i}{\sum_{k=1}^{K} w_{ki} p_{ki}} \tag{12.3}$$

subject to the constraints

$$\frac{\pi_j}{\sum_{k=1}^{K} w_{ki} p_{kj}} \leq 1 \quad j = 1, \ldots, N$$
$$w_{ki} \geq 0 \quad k = 1, \ldots, K$$

This formulation involves a non-linear objective function and non-linear constraints, which is computationally difficult. This problem is often linearized by taking the inverse of the eco-efficiency ratio and solving the associated reciprocal problem (Kuosmanen and Kortelainen 2005; Picazo-Tadeo et al. 2011).

The two constraints in the problem force weights be non-negative and eco-efficiency scores take values between zero and one, that is:

$$EEF_i = \frac{\pi_i}{\sum_{k=1}^{K} w_{ki} p_{ki}} \leq 1, \quad \forall i = 1, \ldots, \mathrm{N} \tag{12.4}$$

The DEA eco-efficiency score which solves this problem for firm i indicates the maximum potential equiproportional reduction in all environmental pressures that could be achieved while maintaining economic value constant, i.e., it corresponds to the ratio OE/OA for a firm operating at point A in Fig. 12.1 and would take the vaue 1 for an eco-efficient firm.

12.3 The SFA Eco-efficiency Model

In this section we introduce our SFA counterpart of the above DEA eco-efficiency model. We first introduce a basic (i.e. homoskedastic) specification of the model in order to focus our attention on the main characteristics of the model and the differences between the SFA and DEA approaches[1]. We then present a heteroskedastic specification of the model that allows us to identify determinants of firms' eco-efficiency in a simple one-stage procedure. Finally, we explain how we obtain the estimates of eco-efficiency for each farm.

12.3.1 Basic Specification

Our SFA approach to modelling eco-efficiency relies on the constraint in Eq. (12.4). If we assume that the environmental pressure weights, w_k, in (12.2) are parameters to be estimated, we can impose that these be positive by reparameterizing them as $w_k = e^{\beta_k}$. As each environmental pressure contributes positively to overall environmental damage, this restriction, which is stated in (12.3), follows naturally. The natural logarithm of Eq. (12.4) can be written as:

$$\ln EFF_i = \ln\left(\frac{\pi_i}{\sum_{k=1}^{K} e^{\beta_k} \cdot p_{ki}}\right) \leq 0 \tag{12.5}$$

The above equation can be rewritten as:

$$\ln(\pi_i) = \ln\left(\sum_{k=1}^{K} e^{\beta_k} \cdot p_{ki}\right) - u_i \tag{12.6}$$

where $u_i = -\ln EFF_i \geq 0$ can now be viewed as a non-negative random term capturing firm i's eco-inefficiency.

[1]A version of this basic homoskedastic model has been presented in Orea and Wall (2015).

Equation (12.6) is a non-linear regression model with a nonpositive disturbance that can be estimated using several techniques, including goal programming, corrected ordiinary least squares (COLS) and modified ordinary least squares (MOLS) —see Kumbhakar and Lovell (2000, Sect. 3.2.1). If we were using a multiplicative aggregation of environmental pressures, we would get a linear (i.e. Cobb-Douglas) regression model where positive parameter values would need to be imposed. Both models are rougly equivalent but the Cobb-Douglas specification would depart from the tradition in the eco-efficiency literature of using linear combinations of environmental pressures.

Regardless of the technique, however, note that in (12.6) we are measuring firms' eco-efficiency relative to a deterministic environmental pressure frontier. This implies that all variation in value added not associated with variation in individual environmental pressures is entirely attributed to eco-inefficiency. In other words, this specification does not make allowance for the effect of random shocks, which might also contribute (positively or negatively) to variations in value added.

As is customary in the SFA literature in production economics, in order to deal with this issue we extend the model in (12.6) by adding a symmetric random noise term, v_i, and a non-zero intercept θ:

$$\ln(\pi_i) = \theta + \ln\left(\sum_{k=1}^{K} e^{\beta_k} \cdot p_{ki}\right) + v_i - u_i \tag{12.7}$$

This model is more complex than a deterministic eco-efficiency frontier model but it is also more realistic as deviations from the frontier due not only to eco-inefficiency but also to uncontrollable or unobservable factors (i.e. random noise) are incorporated. We have also added a non-zero intercept in order to obtain unbiased parameter estimates in case the unobservable factors or measurement errors have a level effect on firms' profit.

The error term in (12.7) thereby comprises two independent parts. The first part, v_i, is a two-sided random noise term, often assumed to be normally distributed with zero mean and constant standard deviation, i.e. $\sigma_v = e^{\gamma}$. The second part, u_i, is a one-sided error term capturing underlying eco-inefficiency that can vary across firms. Following Aigner et al. (1977) it is often assumed to follow a half-normal distribution, which is the truncation (at zero) of a normally-distributed random variable with mean zero. Moreover, these authors also assumed that the variance of the pre-truncated normal variable (hereafter σ_u) is *homoskedastic* and common to all farms, i.e. $\sigma_u = e^{\delta}$. The identification of both random terms in this model (ALS henceforth) relies on the asymmetric and one-sided nature of the distribution of u_i (see Li 1996) If the inefficiency term could take both positive and negative values, it would not be distinguishable from the noise term, v_i.

It should be pointed out that under these distributional assumptions the density function of the composed error term $\varepsilon_i = v_i - u_i$ in (12.7) is the *same* as the well-known density function of a standard normal-half normal frontier model. Following Kumbhakar and Lovell (2000, p. 77), the log likelihood function for a sample of N producers can then be written as:

$$\ln\ L(\theta,\beta,\gamma,\delta) = \frac{N}{2}\ln\left[\sigma_v^2+\sigma_u^2\right] + \sum_{i=1}^{N}\ln\ \Phi\left[-\frac{\varepsilon_i(\theta,\beta)\cdot\sigma_u/\sigma_v}{\left(\sigma_v^2+\sigma_u^2\right)^{1/2}}\right] - \frac{1}{2\left(\sigma_v^2+\sigma_u^2\right)}\sum_{i=1}^{N}\varepsilon_i(\theta,\beta) \tag{12.8}$$

where $\beta = (\beta_1,\ldots,\beta_K)$, and

$$\varepsilon_i(\theta,\beta) = \ln(\pi_i) - \theta - \ln\left(\sum_{k=1}^{K} e^{\beta_k}\cdot p_{ki}\right) \tag{12.9}$$

The likelihood function (12.8) can be maximized with respect to $(\theta,\beta,\gamma,\delta)$ to obtain consistent estimates of all parameters of our eco-efficiency model. The only difference between our SFA *eco-efficiency* model and a traditional SFA *production* model is the computation of the error term $\varepsilon_i(\theta,\beta)$. In a traditional SFA production model, this is a simple linear function of the parameters to be estimated and hence the model can be estimated using standard econometric software, such as Limdep or Stata. In contrast, $\varepsilon_i(\theta,\beta)$ in Eq. (12.9) is a non-linear function of the β parameters. Although the non-linear nature of Eq. (12.9) prevents using the standard commands in Limdep or Stata to estimate our SFA eco-efficiency model, it is relatively straightforward to write the codes to maximize (12.8) and obtain our parameter estimates.

The model in (12.7) can also be estimated using a two-step procedure that combines ML and method of moments (MM) estimators. In the first stage, the intercept θ and the environmental pressure parameters β of Eq. (12.7) can be estimated using a non-linear least squares estimator. In the second step, the aforementioned distributional assumptions regarding the error terms are made to obtain consistent estimates of the parameters describing the variance of v_i and u_i (i.e., γ and δ) conditional on the estimated parameters from the first step. This two-step approach is advocated for various models in Kumbhakar and Lovell (2000). The main advantage of this two-step procedure is that no distributional assumptions are used in the first step. Standard distributional assumptions on v_i and u_i are used only in the second step. In addition, in the first step the error components are allowed to be freely correlated.

An important issue that should be taken into account when using a two-step procedure is that the expectation of the original error term in (12.7) is not zero because u_i is a non-negative random term. This implies that the estimated value of the error term ε_i in Eq. (12.7) should be decomposed as follows:

$$\varepsilon_i = v_i - u_i + E(u_i) \tag{12.10}$$

If u_i follows a half-normal distribution, then $E(u_i) = \sqrt{2/\pi}\cdot\sigma_u$. Thus, the stochastic frontier model in the second stage is:

$$\varepsilon_i = v_i - u_i - \sqrt{2/\pi} \cdot e^{\delta} \tag{12.11}$$

Note that there are no new parameters to be estimated. The parameters γ and δ are estimated by maximizing the likelihood function associated to this (adjusted) error term. As Kumbhakar et al. (2013) have recently pointed out, the stochastic frontier model based on (12.11) can accommodate heteroskedastic inefficiency and noise terms simply by making the variances of σ_u and σ_v functions of some exogenous variables (see, for instance, Wang 2002; Álvarez et al. 2006). This issue is addressed later on.

Before proceeding, it should be pointed out that an alternative two-step approach based only on MM estimatorscan also be used. This empirical strategy relies on the second and third moments of the error term ε_i in Eq. (12.7). This approach takes advantage of the fact that the second moment provides information about both σ_v and σ_u whereas the third moment only provides information about the asymmetric (one-sided) random inefficiency term. Olson et al. (1980) showed using simulation exercises that the choice of estimator (ML vs.MM) depends on the relative value of the variances of both random terms and the sample size. When the sample size is large and the variance of the one-sided error component is small compared to the variance of the noise term, ML outperforms MM. The MM approach has, in addition, some practical problems. It is well known in the stochastic frontier literature, for example, that neglecting heteroskedasticity in either or both of the two random terms causes estimates to be biased. Kumbhakar and Lovell (2000) pointed out that only the ML approach can be used to address this problem. Another practical problem arises in homoskedastic specifications of the model when the implied σ_u becomes sufficiently large to cause $\sigma_v < 0$, which violates the assumptions of econometric theory.

Compared to the DEA eco-efficiency model, our SFA approach will attenuate the effect of outliers and measurement errors in the data on the eco-efficiency scores. Moreover, it is often stressed that the main advantage of DEA over the SFA approach is that it does not require an explicit specification of a functional form for the underlying technology. However, the 'technology' here is a simple index that aggregates all environmental pressures into a unique value. Thus, we would expect that the parametric nature of our SFA approach is not as potentially problematic in an eco-efficiency analysis as it may be in a more general production frontier setting where theses techniques are used to uncover the underlying (and possibly quite complex) relationship between multiple inputs and outputs. Another often-cited advantage of the DEA approach is that it can be used when the number of observations is relatively small. We reiterate, however, that the 'technology' of our SFA model is extremely simple, with few parameters to be estimated, so that the model can be implemented even when the number of observations is not large.

Finally, note that the estimated β parameters have an interesting interpretation in the parametric model. In the expression for eco-efficiency in (12.1), we note that eco-efficiency is constant and equal to 1 along the eco-efficiency frontier.

Differentiating (12.1) in this case with respect to an individual pressure p_k for firm i we obtain:

$$\frac{\partial D_i(p)}{\partial p_k} = \frac{\partial \pi_i}{\partial p_k} \tag{12.12}$$

For any two pressures p_j and p_k, therefore, we have:

$$\frac{\frac{\partial D_i(p)}{\partial p_j}}{\frac{\partial D_i(p)}{\partial p_k}} = \frac{\frac{\partial \pi_i}{\partial p_j}}{\frac{\partial \pi_i}{\partial p_k}} \tag{12.13}$$

From the expression for eco-efficiency in the reparameterized model in (12.5) it is clear that $\frac{\partial D_i(p)}{\partial p_k} = e^{\beta_k}$, so that in this particular case (12.13) becomes:

$$\frac{e^{\beta_j}}{e^{\beta_k}} = \frac{\partial \pi_i / \partial p_j}{\partial \pi_i / \partial p_k} \tag{12.14}$$

Once the β parameters have been estimated, e^{β_k} therefore represents the marginal contribution of pressure p_k to firm i's value added, i.e., it is the monetary loss in value added if pressure p_k were reduced by one unit.

As expression (12.14) represents the *marginal rate of technical substitution of environmental pressures*, it provides valuable information on the possibilities for substitution between pressures. If this marginal rate of substitution took a value of 2, say, we could reduce pressure p_j by two units and increase p_k by one unit without changing economic value added. This also sheds light on the consequences for firms of legislation requiring reductions in individual pressures. Continuing with the previous example, it would be relatively less onerous for the firm to reduce pressure p_k rather than p_j as the fall in value added associated with a reduction in p_k would be only half that which would occur from a reduction in p_j.

12.3.2 Heteroskedastic Specification

Aside from measuring firms' eco-efficiency, we also would like to analyse the *determinants* of eco-efficiency. The concern about the inclusion of contextual variables or z-variables has led to the development of several models using parametric, non-parametric or semi-parametric techniques. For a more detailed review of this topic in SFA and DEA, see Johnson and Kuosmanen (2011, 2012). The inclusion of contextual variables in DEA has been carried out in one, two or even more stages. Ruggiero (1996) and other authors have highlighted that the one-stage model introduced in the seminal paper of Banker and Morey (1986) might lead to bias. To solve this problem, other models using several stages have been developed

in the literature. Ray (1988) was the first to propose a second stage where standard DEA efficiency scores were regressed on a set of contextual variables. This practice was widespread until Simar and Wilson (2007) demonstrated that this procedure is not consistent because the first-stage DEA efficiency estimates are serially correlated. These authors proposed a bootstrap procedure to solve this problem in two stages which has become one of the most-widely used method in DEA to identify inefficiency determinants.

As the inefficiency term in the ALS model has constant variance, our SFA model in (12.7) does not allow the study of the determinants of firms' performance. It might also yield biased estimates of both frontier coefficients and farm-specific eco-inefficiency scores (see Caudill and Ford 1993). To deal with these issues, we could estimate a *heteroskedastic* frontier model that incorporates z-variables into the model as eco-efficiency determinants. The specification of u_i that we consider in this paper is the so-called RSCFG model (see Alvarez et al. 2006), where the z-variables are treated as determinants of the variance of the pre-truncated normal variable. In other words, in our frontier model we assume that

$$\sigma_{ui} = h(z_i) \cdot \sigma_u \tag{12.15}$$

where

$$h(z_i) = e^{\alpha' z_i} \tag{12.16}$$

is a deterministic function of eco-inefficiency covariates, $\alpha = (\alpha_1, \ldots, \alpha_J)$, is a vector of parameters to be estimated, and $z_i = (z_{i1}, \ldots, z_{iJ})$ is a set of J potential determinants of firms' eco-inefficiency. This specification of σ_{ui} nests the *homoskedastic* model as (12.15) colapses into e^{δ} if we assume that $h(z_i) = 1$ or $\alpha = 0$.

The so-called 'scaling property' (Alvarez et al. 2006) is satisfied in this heteroskedastic version of our SFA model in the sense that the inefficiency term in (12.7) can be written as $u_i = h(z_i) \cdot u_i^*$, where $u_i^* \rightarrow N^+(0, e^{\delta})$ is a one-sided random variable that does not depend on any eco-efficiency determinant. The defining feature of models with the scaling property is that firms differ in their mean efficiencies but not in the shape of the distribution of inefficiency. In this model u_i^* can be viewed as a measure of "basic" or "raw" inefficiency that does not depend on any observable determinant of firms' inefficiency.

The log likelihood function of this model is the same as Eq. (12.8), but now σ_{ui} is *heteroskedastic* and varies across farms. The resulting likelihood function should then be maximized with respect to $\theta, \beta, \gamma, \delta$ and α to obtain consistent estimates of all parameters of the model. As both frontier parameters and the coefficients of the eco-inefficiency determinants are simultaneously estimated in one stage, the inclusion of contextual variables in our SFA model is much simpler than in DEA.

12.3.3 Eco-efficiency Scores

We next discuss how we can obtain the estimates of eco-efficiency for each firm once either the homoskedastic or heteroskedastic model has been estimated. In both specifications of the model, the composed error term is simply $\varepsilon_i = v_i - u_i$. Hence, we can follow Jondrow et al. (1982) and use the conditional distribution of u_i given the composed error term ε_i to estimate the asymmetric random term u_i. Both the mean and the mode of the conditional distribution can be used as a point estimate of u_i. However, the conditional expectation $E(u_i|\varepsilon_i)$ is by far the most commonly employed in the stochastic frontier analysis literature (see Kumbhakar and Lovell 2000).

Given our distributional assumptions, the analytical form for $E(u_i|\varepsilon_i)$ can be written as follows:

$$E(u_i|\varepsilon_i) = \bar{\mu}_i + \bar{\sigma}_i \left[\frac{\phi(-\bar{\mu}_i/\bar{\sigma}_i)}{1 - \Phi(-\bar{\mu}_i/\bar{\sigma}_i)} \right] \tag{12.17}$$

where

$$\sigma_i^2 = \sigma_v^2 + h(z_i)^2 \sigma_u^2$$
$$\bar{\mu}_i = \frac{\varepsilon_i h(z_i)^2 \sigma_u^2}{\sigma_i^2}$$
$$\bar{\sigma}_i = \frac{h(z_i)\sigma_u \sigma_v}{\sigma_i}$$

To compute the conditional expectation (12.17) using the heteroskedastic model, we should replace the deterministic function $h(z_i)$ with our estimate of (12.16), while for the homoskedastic model we should assume that $h(z_i) = 1$.

12.4 Data

The data we use come from a survey which formed part of a research project whose objective was to analyse the environmental performance of dairy farmers in the Spanish region of Asturias. Agricultural activity has well-documented adverse effects on the environment, and the increasing concerns among policymakers about environmental sustainability in the sector are reflected in the recent Common Agricultural Policy (CAP) reforms in Europe. Dairy farming, through the use of fertilizers and pesticides in the production of fodder, as well as the emission of greenhouse gases, has negative consequences for land, water, air, biodiversity and the landscape, so it is of interest to see whether there is scope for farmers to reduce

environmental pressures without value added being reduced and identify any farmer characteristics that may influence their environmental performance.

A questionnaire was specifically designed to obtain information on individual pollutants, including nutrients balances and greenhouse gas emissions. These individual pollutants were then aggregated using standard conversion factors into a series of environmental pressures. Questions were included regarding farmers' attitudes towards aspects of environmental management as well as a series of socioeconomic characteristics. The data collected correspond to the year 2010.

A total of 59 farmers responded to the questionnaire and the environmental and socioeconomic data were combined with economic data for these farmers which is gathered annually through a Dairy Cattle Management Program run by the regional government. Given that there were missing values for some of the variables we wished to consider, the final sample comprised 50 farms.

These data were used by Pérez-Urdiales et al. (2015) to measure the farmers' eco-efficiency and relate it to attitudinal and socioeconomic factors. These authors used the two-stage DEA-based bootstrapped truncated regression technique proposed by Simar and Wilson (2007) to estimate eco-efficiency and its determinants, finding evidence of considerable eco-inefficiency. We will use the same variables as Pérez-Urdiales et al. (2015) to estimate eco-efficiency and its determinants using the SFA methods proposed in the previous section, which will permit us to see whether the SFA model yields similar results. We will use the results from Pérez-Urdiales et al. (2015) as a reference for comparison but it should be stressed that the dataset is far from ideal for using a SFA approach. In particular, the number of observations is relatively small and there are several determinants of eco-efficiency whose parameters have to be estimated.

The variables are described in detail in Pérez-Urdiales et al. (2015) but we will briefly discuss them here. For the numerator of the eco-efficiency index, we use the gross margin for our measure of economic value added (*Econvalue*). This is the difference between revenues from milk production (including milk sales and the value of in-farm milk consumption) and direct (variable) costs. These costs include expenditure on feed, the production of forage, expenses relate to the herd, and miscellaneous expenses. Costs related to the production of forage include purchases of seeds, fertilizers and fuel, machine hire and repairs, and casual labour, while herd-related costs include veterinary expenses, milking costs, water and electricity. The environmental pressures comprise nutrients balances and greenhouse gas emissions. The nutrients balances measure the extent to which a farm is releasing

nutrients into the environment, defined as the difference between the inflows and outflows of nutrients. The nutrients balances used are nitrogen (*SurplusN*), phosphorous (*SurplusP*) and potassium (*SurplusK*), all measured in total kilograms. These environmental pressures are constructed using the farm gate balance approach and are calculated as the difference between the nutrient content of farm inputs (purchase of forage, concentrates, mineral fertilizers and animals, legume fixation of nitrogen in the soil and atmospheric deposition) and the nutrient content of outputs from the farm (milk sales and animal sales). The volume of greenhouse gas emissions captures the contribution of the farm to global warming and the dataset contains information on the emissions of carbon dioxide (CO_2), methane (CH_4) and nitrous oxide (N_2O). Each of these greenhouse gases is converted into CO_2 equivalents, so that the variable used is (thousands of) kilos of carbon dioxide released into the atmosphere (CO_2).

The second set of variables are the potential determinants of eco-efficiency, which comprises socioeconomic characteristics and attitudes of farmers. The socioeconomic variables are the age of the farmer (*Age*); the number of hours of specific agricultural training that the farmer received during the year of the sample (*Training*); and a variable capturing the expected future prospects of the farm and which is defined as a dummy variable taking the value 1 if the farmer considered that the farm would continue to be in operation five years later, and 0 otherwise (*Prospects*). As explained in Pérez-Urdiales et al. (2015), eco-efficiency would be expected to be negatively related to age (i.e., older farmers should be less eco-efficient) and positively related to professional training and the expectation that the farm continue.

Three attitudinal variables were constructed from responses to a series of questions on farmers' beliefs regarding their management of nutrients and greenhouse gas emissions as well as their attitudes towards environmental regulation. Thus, on a five-point Likert scale respondents had to state whether they strongly disagree (1), disagree (2), neither agree nor disagree (3), agree (4) or strongly agree (5), with a series of statements regarding their habits and attitudes towards environmental management. The variables $HabitsCO_2$ and *HabitsNutrients* are constructed as dummy variables that take the value 1 if respondents stated that they agreed or strongly agreed that management of grenhouse gases and nutrients was important, and 0 otherwise. The final variable measuring attitudes towards environmental regulation, defined as a dummy variable taking the value 1 if respondants agreed or strongly agreed that environmenatl regulation should be made more restrictive and 0 otherwise (*Regulation*).

Some descriptive statistics of the variables used for measuring eco-efficiency and the determinants of estimated eco-efficiency are presented in Table 12.1.

Table 12.1 Descriptive statistics of variables

Variable	Description	Mean	S. dev.
Econvalue	Value added (€)	77,137	40,423
Environmental pressures			
SurplusN	Nitrogen surplus (kg)	5966	4705
SurplusP	Phosphorous surplus (kg)	2770	2168
SurplusK	Potassium surplus (kg)	2096	1681
CO_2	Greenhouse gases ('000s kg)	427	142
Eco-efficiency determinants			
HabitsCO_2	Attitude towards greenhouse gas management	0.09	0.29
HabitsNutrients	Attitude towards nutrient management	0.77	0.43
Age	Age of head of household	45.98	7.97
Prospects	Continuity of farm	0.98	0.14
Regulation	Attitude towards regulation	0.58	0.50
Training	Hours of specific training in last year	45.14	63.10

12.5 Results

We focus initially on the results from the stochastic frontier models and then on the comparison of these with the DEA results.

Table 12.2 presents estimates from different specifications of the homoskedastic (ALS) and heteroskedastic (RSCFG) stochastic eco-efficiency frontier, with their corresponding eco-efficiency scores presented in Table 12.3. Columns (A) and (B) of Table 12.2 report estimates from the ALS model with all environmental pressures included and it can be seen that all the estimated coefficients on the pressures were highly significant. The parameter δ corresponding to $\ln \sigma_u$ was also highly significant, implying that the frontier specification is appropriate.

Table 12.2 Parameter estimates

Variable	Parameter	ALS-50A		ALS-50B		ALS-50C		ALS-40C		RSCFG-40C	
		(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)
		Estimates	t-stat	Estimates	t-stat	Estimates	t-stat	Estimates	t-stat	Estimates	t-stat
Intercept	θ	−0.15	−1.65	1.42	6.41	2.06	21.79	2.06	21.26	1.96	20.62
SurplusN	β_1	1.59	10.78	0.03	0.02						
SurplusP	β_2	−5.25	−37.11								
SurplusK	β_3	2.24	16.05	0.66	0.31	0.93	7.04	1.17	8.05	1.27	8.63
CO_2	β_4	5.30	42.89	3.73	16.53	3.12	25.13	2.89	20.06	2.91	20.41
ln σ_v	γ	−2.15	−15.19	−2.15	−4.63	−2.26	−3.68	−2.64	−4.67	−2.45	−15.52
ln σ_u	δ	−0.47	−3.41	−0.47	−3.47	−0.45	−2.91	−0.53	−3.92	−0.32	−2.00
HabitsCO_2	α_1									−1.06	−6.71
HabitsNutrients	α_2									−0.21	−1.33
Age	α_3									0.01	1.64
Prospects	α_4									−0.64	−4.06
Regulation	α_5									0.27	1.71
Training	α_6									0.00	−1.35
SurplusN	exp(β_1)	4.90	6.79	1.03	7.15						
SurplusP	exp(β_2)	0.01	7.07								
SurplusK	exp(β_3)	9.37	7.17	1.93	7.10	2.55	7.67	3.22	6.81	3.55	6.80
CO_2	exp(β_4)	200.91	8.09	41.85	8.31	22.74	8.56	17.94	6.99	18.43	7.00
Mean log-likelihood		−0.3999		−0.3999		−0.4039		−0.2867		−0.1744	
Observations		50		50		50		40		40	

Table 12.3 Eco-efficiency scores

Farm	DEA-50	DEA-40	ALS-50A	ALS-50B	ALS-50C	ALS-40C	RSCFG-40C
1	0.615	0.891	0.682	0.683	0.664	0.802	0.878
2	0.562	n.a	0.627	0.627	0.608	n.a	n.a
3	0.233	0.290	0.257	0.257	0.259	0.293	0.305
4	0.407	0.609	0.448	0.448	0.445	0.54	0.577
5	0.906	1.000	0.892	0.892	0.911	0.964	0.961
6	0.698	0.739	0.761	0.761	0.733	0.789	0.818
7	0.690	1.000	0.751	0.751	0.716	0.872	0.905
8	0.540	0.531	0.586	0.587	0.554	0.554	0.578
9	0.790	0.854	0.824	0.824	0.843	0.893	0.906
10	0.307	0.294	0.333	0.334	0.305	0.313	0.327
11	0.809	0.830	0.855	0.855	0.839	0.861	0.879
12	0.675	0.710	0.734	0.734	0.7	0.765	0.797
13	0.617	0.596	0.648	0.648	0.594	0.531	0.539
14	0.797	0.836	0.851	0.851	0.839	0.87	0.886
15	0.465	0.515	0.519	0.519	0.509	0.557	0.581
16	0.238	0.255	0.261	0.261	0.266	0.273	0.281
17	0.702	0.701	0.761	0.762	0.734	0.721	0.74
18	1.000	1.000	0.937	0.937	0.943	0.942	0.937
19	0.566	0.580	0.61	0.61	0.622	0.591	0.606
20	0.515	0.514	0.568	0.568	0.547	0.528	0.543
21	0.559	0.575	0.621	0.621	0.61	0.59	0.603
22	0.567	1.000	0.607	0.607	0.613	0.756	0.806
23	0.844	0.898	0.879	0.879	0.868	0.926	0.932
24	0.298	0.306	0.326	0.326	0.333	0.315	0.323

(continued)

Table 12.3 (continued)

Farm	DEA-50	DEA-40	ALS-50A	ALS-50B	ALS-50C	ALS-40C	RSCFG-40C
25	0.608	0.633	0.661	0.661	0.665	0.652	0.669
26	1.000	n.a	0.937	0.937	0.939	n.a	n.a
27	0.384	n.a	0.432	0.432	0.415	n.a	n.a
28	0.358	0.435	0.404	0.405	0.394	0.445	0.471
29	0.618	0.718	0.681	0.681	0.671	0.751	0.813
30	0.703	n.a	0.746	0.746	0.696	n.a	n.a
31	0.852	0.921	0.848	0.848	0.886	0.935	0.937
32	0.727	0.746	0.784	0.784	0.752	0.787	0.814
33	0.535	0.576	0.582	0.582	0.585	0.61	0.634
34	0.765	n.a	0.823	0.823	0.802	n.a	n.a
35	0.993	n.a	0.935	0.935	0.937	n.a	n.a
36	0.514	0.647	0.54	0.54	0.56	0.643	0.685
37	0.774	0.843	0.818	0.818	0.829	0.888	0.914
38	0.518	0.538	0.547	0.547	0.569	0.554	0.571
39	0.543	0.557	0.594	0.594	0.585	0.577	0.593
40	0.287	0.303	0.321	0.321	0.32	0.318	0.335
41	0.655	0.679	0.722	0.722	0.706	0.704	0.753
42	0.811	n.a	0.837	0.837	0.846	n.a	n.a
43	0.427	0.433	0.467	0.467	0.437	0.466	0.487
44	0.649	n.a	0.685	0.685	0.705	n.a	n.a
45	0.807	0.829	0.849	0.849	0.86	0.84	0.867
46	0.904	0.888	0.906	0.906	0.905	0.871	0.875
47	0.204	n.a	0.234	0.234	0.228	n.a	n.a
48	0.348	n.a	0.39	0.39	0.384	n.a	n.a

(continued)

Table 12.3 (continued)

Farm	DEA-50	DEA-40	ALS-50A	ALS-50B	ALS-50C	ALS-40C	RSCFG-40C
49	0.423	0.450	0.447	0.448	0.399	0.454	0.479
50	0.757	0.755	0.816	0.817	0.8	0.765	0.783
Mean	0.611	0.662	0.647	0.647	0.639	0.663	0.685

As the pressure function parameters β_k enter the eco-efficiency specification exponentially rather than linearly (12.5), in the bottom part of Table 12.2 the exponents of the coefficients are presented. The t-statistics here correspond to the null that e^{β_k} is equal to zero for each of the k pressures, and this is rejected in all cases.

However, focusing on the magnitudes rather than the statistical significance, it can be seen that the marginal contribution of the phosphorous balance to value added is almost negligible. Also, the value of e^{β_k} for potassium is almost twice as large as that of nitrogen. Recalling our discussion of the interpretation of these parameters after Eq. (12.14) above, this implies that potassium contributes twice as much to value added as nitrogen and would therefore be more costly for the farmer to reduce. Similarly, if farmers were required to reduce nitrogen, this could in principle be substituted by potassium: for a given reduction in kilos of nitrogen, farmers could increase their use of potassium by half this number of kilos and maintain the same value added. In this particular application, such substitution could be achieved through changes in the composition of feed, fertilizers, and a change in the composition of forage crops. Reducing phosphorous, on the other hand, would be virtually costless.

In light of the negligible contribution of phosphorous to value added, we rees-timate the ALS model eliminating the phosphorous balance from the pressure function, and the results are presented in columns (C) and (D). The parameters on the nutrients are not significantly differently from 0, implying that the e^{β_k} are not significantly different from 1. Note that the frontier specification is still appropriate and a comparison of the the efficiency scores from the two models in Table 12.3 shows that are practically identical.

We now turn to the heteroskedastic (RSCFG) specification of the stochastic frontier where we incorporate the determinants of eco-efficiency described in the previous section. Some of the farms had missing values for one or more of these determinants, and after eliminating these observations we were left with 40 farms with complete information. When estimating the model for these 40 observations with all nutrients balances included, it did not converge. We then eliminated the phosphorous balance as we had done in columns (C) and (D) for the homoskedastic (ALS) specification, but the model still did not converge. Following our earlier strategy of eliminating the nutrient balance with the lowest marginal contribution, from column (A) we see that the nitrogen balance has a far lower marginal contribution than the potassium balance. We therefore specified the model without the nitogen balance, keeping only the potassium balance and grenhouse gas emissions as pressures. With this specification the model converged successfully and the results are reported in columns (I) and (J). The homoskedastic specification with the potassium and greenhouse gases as the only pressures for both the complete sample of 50 observations (ALS-50C) and the reduced sample of 40 observations (ALS-40C) are reported in Columns (E)-(H), and a comparison of these estimates reveals that the coefficients on the pressures change very little across the three models.

To compare the eco-efficiencies estimated by DEA and SFA, a scatterplot of the DEA efficiency scores and the efficiency estimates from the SFA model is presented in Fig. 12.2. These DEA scores are based on a simple DEA calculation as opposed to the bootstrapped DEA scores reported in Pérez-Urdiales et al. (2015). As can be seen, the eco-efficiencies are almost identical. Also plotted on Fig. 12.2 is the regression line from the regression of the SFA estimates on the DEA scores, where the R^2 is 0.9805. The Spearman Rank Correlation Coefficient (Spearman's rho) was 0.996, showing that the models yielded virtually identical rankings of eco-efficiency levels. Even when the reduced sample of 40 observations is used, the eco-efficiencies are again very similar, with almost identical mean values and a Spearman Rank Correlation Coefficient of 0.959.

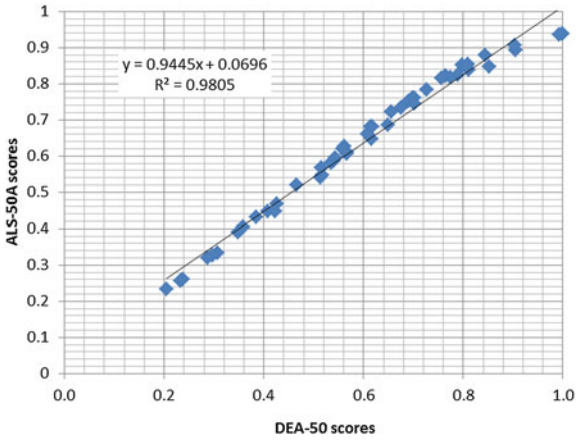


Fig. 12.2 Comparison of Eco-efficiency scores

While the raw eco-efficiency scores between DEA and SFA are very similar, the questions remains as to whether the models yield similar results with regard to the determinants of eco-efficiency. The estimates of the efficiency determinants from the SFA model from Table 12.2 are presented in Table 12.4 alongside the parameter estimates reproduced from Pérez-Urdiales et al. (2015). While all the determinants in Pérez-Urdiales et al. (2015) were found to be significant at the 95% level, only two of the determinants—*HabitsCO*$_2$ and *Prospects*—are significant at this level in the heteroskadastic SFA model (though two other variables—*Age* and *Regulation* - were significant at the 90% level). Notably, however, the SFA model yields exactly the same signs on the eco-efficiency determinants as the bootstrapped DEA-based truncated regression used by Pérez-Urdiales et al. (2015).

Table 12.4 Estimated coefficients and significance of eco-efficiency determinants

	SFA		DEA	
Variable	Estimated parameter	Significant at 95% level?	Estimated parameter	Significant at 95% level?
HabitsCO_2	−1.060	Yes	−0.689	Yes
HabitsNutrients	−0.210	No	−0.231	Yes
Age	0.011	No	0.008	Yes
Prospects	−0.641	Yes	−2.144	Yes
Regulation	0.270	No	0.230	Yes
Training	−0.004	No	−0.002	Yes
Intercept	−0.317	Yes	0.161	Yes

The SFA results come from Table 12.2. The DEA results are obtained from Pérez-Urdiales et al. (2015, Table 12.3)

12.6 Conclusions

Measurement of eco-efficiency has been carried out exclusively using non-parametric DEA techniques in the literature to date. In the present work we have proposed using a (parametric) stochastic frontier analysis (SFA) approach. While such models are highly non-linear when estimating eco-inefficiency, in an empirical application we find that such an approach is feasible even when the sample size is relatively small and determinants of eco-inefficiency - which increases the number of parameters to be estimated - are incorporated. Using data from a sample of 50 Spanish dairy farms previously used by Pérez-Urdiales et al. (2015), we begin by estimating a stochastic frontier model without eco-efficiency determinants, and find that our model yields virtually identical eco-efficiency scores to those calculated by DEA. Estimating eco-efficiency without determinants involves relatively few parameters, so sample size should not be a major obstacle to using SFA. Our results corroborate this.

We then estimated a heteroskedastic SFA model which incorporated determinants of eco-inefficiency. We use the same determinants used by Pérez-Urdiales et al. (2015), who carried out their analysis applying bootstrapped truncated regression techniques. As extra parameters have to be estimated, the small sample size became more of an issue for the stochastic frontier model. Indeed, in order for the model to converge we had to use fewer environmental pressures in our application than Pérez-Urdiales et al. (2015). Encouragingly, however, we found the exact same signs on the determinants of eco-efficiency as those found by Pérez-Urdiales et al. (2015). Thus, even with a small sample size and multiple determinants of eco-inefficiency, the stochastic frontier model yields similar conclusions to those obtained by truncated regression techniques based on DEA estimates of eco-efficiency.

Using stochastic frontier models for eco-efficiency measurement has some advantages over the bootstrapped truncated regression techniques that have been

employed in the litrature to date. In particular, the stochastic frontier model can be carried out in one stage and the coefficients on the environmental pressures ('technology' parameters) have interesting interpretetations which shed light on the contribution of these pressures to firm economic value added. The estimated coefficients also uncover potentially useful information on the substitutability between environmental pressures. As such, we advocate the use of SFA for measuring eco-efficiency as a complement to or substitute for DEA-based approaches. When sample size is small and we wish to incorporate determinants of eco-efficiency, the DEA-based truncated regression techniques may permit more environmental pressures to be included in the analysis. However, with larger sample sizes, we would expect this advantage to disappear and the stochastic frontier models can provide extra valuable information for producers and policy-makers, particularly with regard to substitutability between pressures.

References

Aigner D, Lovell K, Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. J Econometrics 6:21–37

Alvarez A, Amsler C, Orea L, Schmidt P (2006) Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics. J Prod Anal 25:201–212

Banker RD, Morey R (1986) Efficiency analysis for exogenously fixed inputs and outputs. Oper Res 34:513–521

Caudill SB, Ford JM (1993) Biases in frontier estimation due to heteroscedasticity. Econ Lett 41:17–20

Coelli T, Lauwers L, Huylenbroeck GV (2007) Environmental efficiency measurement and the materials balance condition. J Prod Anal 28(1):3–12

Johnson AL, Kuosmanen T (2011) One-stage estimation of the effects of operational conditions and practices on productive performance: Asymptotically normal and efficient, root-n consistent StoNEZD method. J Prod Anal 36(2):219–230

Johnson AL, Kuosmanen T (2012) One-stage and two-stage DEA estimation of the effects of contextual variables. Eur J Oper Res 220(2):559–570

Jondrow J, Lovell K, Materov I, Schmidt P (1982) On the estimation of technical inefficiency in the stochastic frontier production function model. J Econometrics 19:233–238

Kumbhakar SC, Lovell CAK (2000) Stochastic frontier analysis. Cambridge University Press, Cambridge

Kumbhakar SC, Asche F, Tveteras R (2013) Estimation and decomposition of inefficiency when producers maximize return to the outlay: an application to Norwegian fishing trawlers. J Prod Anal 40:307–321

Kuosmanen T, Kortelainen M (2005) Measuring eco-efficiency of production with data envelopment analysis. J Ind Ecol 9(4):59–72

Lauwers L (2009) Justifying the incorporation of the materials balance principle into frontier-based eco-efficiency models. Ecol Econ 68:1605–1614

Li Qi (1996) Estimating a stochastic production frontier when the adjusted error is symmetric. Econ Lett 52(3):221–228

OECD (1998) Eco-Efficiency. Organization for Economic Co-operation and Development, OECD, Paris

Olson JA, Schmidt P, Waldman DM (1980) A monte carlo study of estimators of stochastic frontier production functions. J Econometrics 13:67–82

Orea L, Wall A (2015) A parametric frontier model for measuring eco-efficiency, Efficiency Series Paper ESP 02/2015, University of Oviedo

Oude Lansink A, Wall A (2014) Frontier models for evaluating environmental efficiency: an overview. Econ Bus Lett 3(1):43–50

Pérez-Urdiales M, Lansink Oude, Wall A (2015) Eco-efficiency among dairy farmers: the importance of socio-economic characteristics and farmer attitudes. Environ Resource Econ. doi:10.1007/s10640-015-9885-1

Picazo-Tadeo AJ, Reig-Martínez E, Gómez-Limón J (2011) Assessing farming eco-efficiency: a data envelopment analysis approach. J Environ Manage 92(4):1154–1164

Picazo-Tadeo AJ, Beltrán-Esteve M, Gómez-Limón J (2012) Assessing eco-efficiency with directional distance functions. Eur J Oper Res 220:798–809

Ray SC (1988) Data envelopment analysis, nondiscretionary inputs and efficiency: An alternative interpretation. Socio-Economic Plan Sci 22(4):167–176

Ruggiero J (1996) On the measurement of technical efficiency in the public sector. Eur J Oper Res 90:553–565

Schmidheiney S (1993) Changing course: A global business perspective on development and the environment, Technical report. MIT Press, Cambridge

Simar L, Wilson P (2007) Estimation and inference in two-stage, semiparametric models of production processes. J Econometrics 136:31–64

Tyteca D (1996) On the measurement of the environmental performance of farms - a literature review and a productive efficiency perspective. J Environ Manage 68:281–308

Wang H-J (2002) Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model. J Prod Anal 18:241–253

Chapter 13
Bankruptcy Prediction of Companies in the Retail-Apparel Industry Using Data Envelopment Analysis

Angela Tran Kingyens, Joseph C. Paradi and Fai Tam

Abstract Since 2008, the world has gone through a significant recession. This crisis has prompted many small businesses and large corporations to file for bankruptcy, which has grave global social implications. While the markets have recovered much of the lost ground by now, there is still great opportunity to learn about all the possible factors of this recession. We develop a model using DEA to predict the likelihood of failure of US companies in the retail-apparel industry based on information available from annual reports—financial statements and their corresponding Notes, Management's Discussion and Analysis, and Auditor's Report. It was hypothesized that variables that reflect managerial decision-making and economic factors would enhance the predictive power of current mathematical models that consider financial data exclusively. This is an effective prediction tool, separating companies with a high risk of bankruptcy from those that were healthy, with a reliable accuracy of 80%—an improvement over the widely-used Altman bankruptcy model having 70, 58 and 50% accuracy when predicting cases today, from one year back and from two years back, respectively.

13.1 Introduction

Bankruptcy is a legally declared financial inability of an individual or organization to meet its debt obligations. In the US, nearly 1.59 million filings for bankruptcy protection were made in 2010, an increase of 8% from 2009 and 43% from 2008 (US Courts, Bankruptcy Statistics 2010). The effect of bankruptcy is two-fold. Creditors lose at least part of their interest payments and principal investment, while

A.T. Kingyens · J.C. Paradi (✉) · F. Tam
University of Toronto, Toronto, Canada
e-mail: paradi@mie.utoronto.ca
URL: http://versionone.vc/about-us/

A.T. Kingyens
versionone, Vancouver, Canada
e-mail: angela@versionone.vc
URL: http://versionone.vc/about-us/

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_13

business owners are subject to unpleasant legal, financial and personal consequences. For example, bankruptcy may harm the owner's reputation, damage existing relationships and make it difficult to obtain credit in the future. Thus there are economic and social incentives for predicting bankruptcy. Creditors could better assess risk and price interest accordingly. For owners, prediction would translate to prevention, buying time for them to secure more financing and avoid potential demise. Investors could use the information to invest in healthy companies and short those at risk.

Most of the prevalent mathematical models used for bankruptcy prediction rely on financial data and ratios, ignoring managerial, market and economic factors. Also, while these techniques generally make dichotomous (i.e. survive or fail) predictions, they do not provide owners with strategies to improve their operations when bankruptcy is looming. Therefore the objective of this paper is to develop a bankruptcy prediction model based on Data Envelopment Analysis (DEA) that will assess the likelihood of bankruptcy and suggest preventive measures. The model considers data from financial statements and their accompanying Notes (which provide hints on managerial decision-making) as well as market and economic influences, as it is hypothesized their inclusion will enhance predictive power. The performance of this new model is compared with that of the industry standard *Altman Z-score*, which had accuracy rates of 70, 58 and 50% predicting bankruptcies on the data studied today, one year, and two years back, respectively. Although the model is adaptable to other industries, the data used was from the American retail-apparel industry.

The remainder of this paper is organized as follows. Section 13.2 gives the background and a literature review of prior work on bankruptcy prediction is presented, as well a basic description of DEA, the basis of the developed model. Section 13.3 describes the US retail-apparel industry and financial statement data utilized. Section 13.4 presents the developed methodology. Section 13.5 details the formulations and results of the individual financial statement DEA models used as first-stage metrics. Section 13.6 presents the results of the second-stage DEA model. Discussion of the results is given in Sects. 13.7 and 13.8 offers conclusions.

13.2 Background

Bankruptcy is a legally declared financial inability of an organization to meet its debt obligations. It is distinguished from insolvency which is a financial state where a company's liabilities exceed its assets, or it is unable to meet debt payments as they become due. Causes of bankruptcy include: financial mismanagement, poor location, competition, difficulties with cash flow, loss of capitalization, high debt, lack of planning, tax burdens and regulations, loss of a key person, lack of technology, poor record keeping, personal issues, and natural disaster or accident leading to high insurance (Bradley and Cowdery 2004).

Bankruptcy prediction has been a prolific field of study, and researchers have considered various factors such as financial ratios, geographic location, type of

industry and competition. To date, there are over 200 journal articles reporting on bankruptcy prediction using common financial ratios alone. Ravi Kumar and Ravi (2007) conducted an extensive review of nearly 120 papers, grouping them by methodology into nine categories: statistical techniques, neural networks, case-based reasoning, decision trees, operations research, evolutionary approaches, rough set based techniques, other techniques subsuming fuzzy logic and support vector machine and isotonic separation, and combinations of multiple techniques. Statistical techniques are amongst the most popular and serve as the benchmark for the developed model, and operations research is the basis for the model presented here.

One of the standard bankruptcy prediction models is the Altman Z-score (Altman 1968), which was derived by applying multiple discriminant analysis to common financial ratios, yielding a single score from a linear combination of these ratios. The range of Z-scores was broken down into three zones: high scores representing the safe (non-bankrupt) zone, and low scores the distress (bankrupt) zone. No prediction was made for firms with intermediate scores (the grey zone). Despite the simplicity of the Z-score, it has been found to work well in a variety of circumstances. The original model was developed on a 50–50 dataset of bankrupt and non-bankrupt public manufacturing companies, and was able to predict bankruptcies up to 2 years beforehand with overall error rates of 5 and 17% in years 1 and 2 respectively.

Two further models were proposed for private firms and public non-manufacturing firms (Hanson 2003). The public non-manufacturing model is given by:

$$Z' = 6.56 \times \frac{\text{Working Capital}}{\text{Total Assets}} + 3.26 \times \frac{\text{Retained Earnings}}{\text{Total Assets}} + 6.72 \times \frac{\text{Earnings Before Interest and Taxes}}{\text{Total Assets}} + 1.05 \times \frac{\text{Market Value of Equity}}{\text{Market Value of Total Liabilities}}$$

$Z' \geq 2.6$ represents the safe zone, and $Z' \leq 1.1$ is the distress zone.

Another statistical model is the Ohlson model (Ohlson 1980), based the maximum likelihood estimate of the conditional logit model to predict bankruptcy. It was derived using data on publicly-listed industrial firms from 1970 to 1976, representing a sample of 105 bankrupt and 2068 non-bankrupt firms, which was more representative of reality than the 50–50 split used to develop the Altman Z-score

A comparison of the two models (see Ohlson 1980 for the formulation) shows the Z-score considers inefficient use of assets as a key driver for bankruptcy, while the Ohlson model incorporates firm size and considers large firms less likely to go bankrupt.

Beaver (1966), Wilcox (1971), Deakin (1972), Edmister (1972), Blum (1974), Libby (1975) and Moyer (1977) also considered statistical methods (e.g. quadratic discriminant analysis, logical regression and factor analysis) for bankruptcy prediction. Some issues identified from these studies were that the proportion of bankrupt firms in the economy is much smaller than non-bankrupt firms. However,

prediction was generally better with a higher proportion of bankrupt firms in the data, suggesting that using representative data samples is important to avoid bias. As industries have different structures and environments, it is important to develop models for specific industries, and they should be normalized for firm size. Companies are subject to market and economic forces which change over time. Bankruptcy prediction models may not be stationary and should be periodically updated, including reflecting new financing vehicles that become available.

Data Envelopment Analysis (DEA) is a non-parametric fractional linear programming technique that can rank and compare the relative performance of DMUs operating under comparable conditions. It is particularly useful in cases where DMUs use multiple inputs to produce multiple outputs. DEA arrives at a score for each DMU relative to the entire data sample (production possibility set) as the ratio of a combined virtual output to virtual input. The frontier of empirically efficient DMUs provides a benchmark or target for the changes required in inputs and outputs to render an inefficient DMU efficient.

The basic DEA model choices include the returns to scale assumption (variable or constant), orientation (input, output or non-oriented), and model type—c.f. Cooper et al. (2007) for a detailed treatment of these, including respective model formulations. The shape of the efficient frontier is not affected by the choice of orientation for any of the basic models; however, the projection of inefficient DMUs onto the frontier can differ markedly.

The work presented herein employs the non-oriented form of the slack-based measure (SBM) DEA model. The linearized form of this model is expressed as (Tone 2001):

$$
\begin{aligned}
\text{Minimize} \quad & \tau = t - \frac{1}{m}\sum_{i=1}^{m} S_i^- / x_{io} \\
\text{Subject to} \quad & 1 = t + \frac{1}{s}\sum_{i=1}^{s} S_r^+ / y_{ro} \\
& \sum_{j=1}^{n} x_{ij}\Lambda_j + S_i^- = t x_{io} \quad (i = 1, \ldots, m) \\
& \sum_{j=1}^{n} y_{rj}\Lambda_j - S_r^+ = t y_{ro} \quad (r = 1, \ldots, s) \\
& \Lambda_j \geq 0 \quad (j = 1, \ldots, n) \\
& S_i^- \geq 0 \quad (i = 1, \ldots, m) \\
& S_r^+ \geq 0 \quad (r = 1, \ldots, s) \\
& t > 0
\end{aligned}
\qquad (13.1)
$$

where τ is efficiency score (ρ) for DMU_o, t is a scaling parameter introduced to linearize the program, m is the number of input variables, s is the number of outputs, x_{ij} is the amount of the ith input to DMU_j, y_{rj} is the amount of the rth

output from DMU_j, S_i^- is the scaled slack on input i, S_r^+ is the scaled slack on output r, and Λ_j is the scaled intensity variables representing the weight of each DMU_j in the benchmark reference for DMU_o. The optimal solution for DMU_o is defined by: $\rho^* = \tau^*$, $\boldsymbol{\lambda}^* = \boldsymbol{\Lambda}^*/t^*$,$\boldsymbol{s}^{-*} = \boldsymbol{S}^{-*}/t^*$, $\boldsymbol{s}^{+*} = \boldsymbol{S}^{+*}/t^*$.

The VRS formulation has the additional constraint that the sum of Λ_j = t.

The SBM model cannot have negative data, and any zero outputs need to be replaced with a very small positive constant to prevent division by zero.[1] The SBM model can be deemed to measure both technical and mix efficiencies as it permits the proportions or mixes of inputs and outputs to vary.

There are a couple of limitations to DEA worth mentioning as they are relevant to the work presented. DEA cannot accurately model small sample sizes, and yielding in high efficiency scores and a high proportion of efficient DMUs. The rough rule of thumb for minimum sample size is

$$n \geq \max\{m \times s, 3(m+s)\}, \tag{13.2}$$

where n, m and s are the number of DMUs, inputs and outputs, respectively (Banker et al. 1989). It is also problematic to use inputs or outputs in ratio form. DEA would evaluate the ratio variable for a possible production as the weighted sum of individual ratios, whereas the correct calculation would be the ratio of the weighted sums of the numerator and denominator. This can lead to a misspecification of the efficient frontier, inaccurate scores and distorted projections (Hollingsworth et al. 2003;Emrouznejad and Amin 2007).

DEA's advantages and the intuitive relationship between inefficiency and failure have led to many studies employing DEA to predict failures, primarily in banks. Ravi Kumar and Ravi (2007) provide a comprehensive survey. Barr et al. (1993, 1994) incorporated the international bank ratings system CAMELS (Capital adequacy, Asset quality, Management quality, Earnings, Liquidity, Sensitivity to market risk) with DEA to predict failures for 930 banks over 5 years. Kao et al. (2004) incorporated financial forecasts from general managers into DEA to predict bankruptcy in a sample of 24 Taiwanese banks. The results of both studies were promising.

Extending beyond banks to more general companies, Cielen et al. (2004) used 11 financial ratios in a DEA model to predict bankruptcy with type I and type II errors of 21 and 24%, respectively, in a sample of 276 non-bankrupt and 90 bankrupt firms between 1994 and 1996. Premachandra et al. (2011) used an additive DEA model on over 1000 US firms in a variety of industries from 1991 to 2004, which was relatively weaker at predicting failures relative to correctly classifying healthy firms. Xu et al. (2009) showed that using DEA scores as a variable (representing operational efficiency) in other bankruptcy prediction methodologies,

[1]As the constraints in (1) prevent any increase in inputs, the projection for a DMU with a zero input will also use none of that input, so 0/0 for inputs can be interpreted as zero in the objective function, i.e. no improvement.

i.e. support vector machines, logistic regression and MDA, improved prediction accuracy. However, their data sample had a non-representative 50–50 proportion of bankrupt and non-bankrupt firms.

New types of DEA models have been created for bankruptcy studies. Paradi et al. (2004) analyzed manufacturing companies with their worst practice DEA model, i.e. a model formulated to identify distressed firms as "efficient". This model performed well in the initial data sample that had equal numbers of bankrupt and non-bankrupt firms; however, it did not perform as well on a larger, more representative dataset. Sueyoshi (1998), and Sueyoshi and Goto (2009) developed DEA-DA, combining an additive DEA model with discriminant analysis, and applied it to bankruptcy analysis. It was designed to find overlaps between two groups (i.e. bankrupt and healthy) in a first-stage, and determine a piecewise linear classification function to separate the groups in a second-stage. In the dataset studies consisting of 951 non-bankrupt DMUs and 50 bankrupt, DEA alone misclassified bankrupts due to the small proportion of bankrupts, while DEA-DA was able to deal with the data imbalance by controlling the importance of the two groups, as determined by size.

13.3 US Retail-Apparel Industry—Data Collection and Exploration

The US retail-apparel industry (classification: SIC Division G, Major Group 56) was chosen for the study because the industry is sufficiently competitive such that turnover of firms is not uncommon, and its environment was relatively consistent over the time period considered, other than the effects of inflation and recession. The industry is characterized by low capital investment requirements and barriers to entry, resulting in high competition levels and low profit margins. Thus, the strategic focus to increase returns in these industries is to increase asset turnover (Stickney et al. 2006). In addition, its firms were comparable, being of similar size, and operating with similar business models in the same country. A total of 85 firms that were active and publicly traded from 1996 to 2009 were considered, 24 of which had declared bankruptcy at least once during this period. The most common reasons for the bankruptcy filings were debt burden, change in management, fraud and recession.

Each combination of a firm and year of operation was considered a separate DMU; this is a common DEA practice when analysis is across time periods. A DMU was considered as bankrupt up to three years prior to the year in which a firm filed for bankruptcy; otherwise, it was considered active. In total, there was 701 DMUs, 50 of which were bankrupt. The financial information for each firm was taken from the spreadsheets in the Capital IQ database. Common templates for financial statements were created using the Financial Statement Analysis Package (Stickney et al. 2006) to address any issues arising from inconsistencies in data labelling between companies. These templates were populated via Visual Basic

code, verified via balance checks and matching important totals with the original statements, and any discrepancies were manually addressed.

Financial variables from the three major financial statements, i.e. the balance sheet (BS), income statement (IS) and cash flow statement (CFS), were considered for inclusion into the model development. In the final model, no variables were taken from the CFS. Cost of goods sold (COGS); income tax expense (ITE); net interest expense (IE); sales, general and administrative expenses (SGA); net income (NI) and revenue were used from the income statement. Variables taken from the balance sheet were: accounts receivable (AR); current assets (CA); 'goodwill'; inventory; marketable securities (MS); net property, plant and equipment (PPE); long-term investment in securities (LTIS); total assets; cash; retained earnings (RE); shareholders' equity (SE); accounts payable (AP); current liabilities (CL); long-term debt (LTD); current maturities of LTD (CM); notes payable and short-term debt (NPSTD); and total liabilities (TL).

The strongest correlations in the BS and IS were found between the larger, i.e. total, amounts. Between the variables from the two statements, revenue, COGS, and SGA (which were themselves highly correlated) were found be strongly correlated with asset and liability items. The variables with strong correlations (above 0.85 and significant at the 0.01 level) are detailed in Table 13.1.

Common profitability and solvency ratios were computed and compared between the active and bankrupt DMUs (c.f. Gill 1994 for the definition of common financial variables and ratios employed throughout this paper). For each firm, the median values of its ratios were taken across all the years for which there was data. The firms were divided into two states: bankrupt if they had ever declared bankruptcy in the 1996–2009 time period, and active otherwise; and averages of these medians values were taken over all firms in a particular state. These ratio values are summarized in Table 13.2.

Profit margin and return on assets were found to be higher for active firms. Total asset turnover was not a significant predictor of failures, contrary to expectations that it would be the most important component of profitability for a highly competitive industry (Stickney et al. 2006). Active firms had higher accounts receivables and inventory turnovers, and higher return on equity. They also had better short-term liquidity ratios and more days of revenue as cash on hand, higher interest coverage, and lower ratios of liabilities to assets and equity.

The annual reports for the firms studied, with the exception of BSST and CSS, were available from EDGAR, the SEC and/or company websites. From the abundant information contained in the Notes, MD&A and Auditor's Report sections, six types were chosen to be extracted based on their satisfying four characteristics: relevance to the retail-apparel industry, being commonly reported across all companies, ease of extraction from the annual report, and ability to be translated into a useful (e.g. binary) scale. These were:

1. *Significant Related-Party Transactions*, in the form of leases or loans to the executive, etc.;
2. *Auditor's Opinion*, either unqualified or qualified;

Table 13.1 Highly correlated accounts on the balance sheet and income statement

Accounts	Correlation	Accounts	Correlation
Inv, CA	0.95	AP, TL	0.91
Inv, PPE	0.87	CM, CL	0.94
Inv, TA	0.94	CM, TL	0.93
Inv, AP	0.94	CL, TL	0.97
Inv, CM	0.87	CL, SE	0.86
Inv, CL	0.94	RE, SE	0.86
Inv, TL	0.92	Revenue, COGS	0.99
Inv, SE	0.86	Revenue, SG&A	0.98
CA, PPE	0.91	COGS, SG&A	0.95
CA, TA	0.98	ITE, Net Income	0.87
CA, AP	0.93	Inv, Revenue	0.86
CA, CM	0.90	Inventories, COGS	0.86
CA, CL	0.95	Inventories, SG&A	0.85
CA, TL	0.93	CA, Revenue	0.89
CA, SE	0.93	CA, COGS	0.87
PPE, TA	0.93	CA, SG&A	0.88
PPE, AP	0.87	PPE, SG&A	0.85
PPE, CM	0.87	TA, Revenue	0.89
PPE, CL	0.89	Total Assets, COGS	0.87
PPE, TL	0.88	Total Assets, SG&A	0.89
PPE, SE	0.88	AP, Revenue	0.86
TA, AP	0.92	AP, COGS	0.86
TA, CM	0.91	CL, Revenue	0.89
TA, CL	0.95	CL, COGS	0.88
TA, TL	0.95	CL, SG&A	0.88
TA, SE	0.94	TL, Revenue	0.87
AP, CM	0.87	TL, COGS	0.85
AP, CL	0.96	TL, SG&A	0.87

3. *Independent Auditing Company*;
4. *Legal proceedings*, such as those arising in the normal course of business that do not have a material adverse effect on the company, litigations that lead to significant payouts and filing of Chap. 11 bankruptcy;
5. *Name of Chairman*, *Name of Chief Executive Officer*, and *Name of Chief Financial Officer*; and,
6. *Retirement plans*, where if applicable, employees are eligible to participate in the company's 401(k) plan or there is a specific company-sponsored pension program.

Table 13.2 Average median ratio values by state

Ratio Type	Ratio	Active	Bankrupt	All
Profitability	TA turnover	2.4 ± 0.9	2.7 ± 0.9	2.4 ± 0.9
	Profit margin	4.2 ± 3.0%	−3.1 ± 6.0%	2.1 ± 5.5%
	ROA	10.6 ± 6.0%	−3.3 ± 14.7%	6.9 ± 10.9%
	AR turnover	150 ± 220	96 ± 95	147 ± 210
	Inventory turnover	100 ± 290	15 ± 20	94 ± 280
	ROE	15.9 ± 13.7%	−7.3 ± 20.9%	9.4 ± 19.1%
Liquidity	Current ratio	2.5 ± 1.1	1.8 ± 1.0	2.4 ± 1.1
	Quick ratio	1.1 ± 1.0	0.6 ± 0.6	1.1 ± 0.9
	CFO to CL ratio	0.6 ± 0.6	0.2 ± 0.5	0.6 ± 0.6
	Days revenue in cash	35 ± 66	12 ± 15	33 ± 64
Solvency	Interest coverage	41.7 ± 79.1	9.1 ± 45.4	38.6 ± 77.1
	Liabilities to assets	0.4 ± 0.2	0.7 ± 0.3	0.5 ± 0.2
	Liabilities to SE	0.9 ± 1.2	1.3 ± 3.0	0.9 ± 1.4
Market performance	Return	0.013 ± 0.055	−0.019 ± 0.099	0.011 ± 0.059
	Volatility	0.16 ± 0.08	0.26 ± 0.15	0.17 ± 0.09
	Beta	1.16 ± 1.45	1.63 ± 3.11	1.19 ± 1.61

This qualitative information was translated into variables representing management decision making (MDM), detailed in Table 13.3. Numerical values were given for each category, which were mostly binary or counts. Legal proceedings were assigned a range of 0–25 based on the frequency of occurrence of each category in the data.

As the financial and MDM data were collected from audited financial statement, missing data and errors were not significant concerns. The DMU (firm-year pair) was omitted if the annual report could not be found. Extreme year-to-year changes were scaled down to 20 times the average change in magnitude. This was done to keep the DMU and recognize the extreme change it represented, without skewing the rest of the data.

Seventeen variables of economic data compiled by US agencies were considered for the model. There were chosen on the bases of having a direct effect on the retail-apparel industry and being mostly weakly correlated (magnitude < 0.4) between each other. Table 13.4 lists these variables, their units of measure, and their timing (leading/coincident/lagging) and relation (pro- or countercyclical) to the business cycle.

The influence of the market was proxied by the returns of the NYSE and NASDAQ composite indices. Of the 85 firms in the data, 38, 35 and 17 traded on the NASDAQ, NYSE and over the counter (OTC) markets respectively. Active

Table 13.3 Managerial decision-making (MDM) variables

Variable	Outcome	Value assigned
Significant related party transactions	None	0
	Yes	1
Auditor's opinion	Unqualified	0
	Qualified	1
Legal proceedings	None	0
	Insignificant	2
	Significant	10
	Going concern	20
	Bankruptcy Filing	25
Retirement plan	None	0
	Yes	1
Auditor change (change in auditor company)	None	0
	Yes	1
Turnover of management (change in either chairman, chief executive officer or chief financial officer)	None	0
	Yes	1, 2 or 3 (depending on how many changed in that year)

firms had higher returns, lower volatility and betas (see Table 13.2). Table 13.5 shows that the correlations between market return, firm returns, volatilities and betas were weak, suggesting that the variables should be individually considered.

The various variables detailed in this section (financial, MDM, economic and market) were found to be poor predictors of bankruptcy when used individually, and exhibited weak mutual correlation. This reinforced the motivation to combine them into a single model or methodology to capture more and different dimensions of company health.

13.4 Methodology

The objective of this research is to develop a (DEA)-based model which predicts the likelihood of failure of American retail-apparel companies (although the developed methodology can be adapted to other industries) and suggests preventative measures based upon the results of its analysis. It is hypothesized that supplementing the data available from financial statements with their accompanying Notes (which provide hints on managerial decision-making), as well as market and economic influences, will enhance the predictive power of mathematical models that consider financial data exclusively. A summary of how DEA fits into the bankruptcy prediction landscape, along with some current relevant limitations, is

Table 13.4 Economic factors considered

Factor		Measured in	Analysis in	Timing and relation to business cycle
General economic factors				
E1	GDP rate	$B	%	Coincident, procyclical
E2	Debt as % of GDP	%	Same	Coincident, countercyclical
E3	Inflation	%	Same	Coincident, procyclical
E4	Interest rate	%	Same	Coincident, procyclical
E5	Unemployment rate	%	Same	Lagged, countercyclical
Apparel factors				
E6	Personal consumption expenditures: clothing and footwear	$B	%	Coincident, procyclical
E7	GDP: clothing and footwear	$B	%	Coincident, procyclical
E8	CPI: apparel	Index (1982)	%	Coincident, procyclical
E9	Industrial production: clothing	Index (2007)	%	Coincident, procyclical
E10	Apparel unit labor cost = (total labour compensation/hours)/productivity	%	Same	Coincident, procyclical
E11	Apparel labor productivity = output/hours	%	Same	Coincident, procyclical
E12	Apparel imports	$	%	Coincident, procyclical
E13	Apparel exports	$	%	Coincident, countercyclical
Other factors				
E14	New privately owned housing units started	1000 s	Same	Leading, procyclical
E15	Median number of months for a sale	Months	Same	Leading, countercyclical
E16	Oil price	$/bbl	Same	Coincident, countercyclical
E17	Cotton price	cent/lb	Same	Coincident, countercyclical

presented in Fig. 13.1. This work aims to address some of these limitations by introducing novel measures incorporating contributions from the fields of finance, accounting, and operations research.

DEA was chosen as the underlying method for the model, as it can be used to determine the state of each company within a sample. The non-oriented, VRS form

Table 13.5 Correlations between composite indices and stock performance

	Return		Volatility		Beta	
	Active	Bankrupt	Active	Bankrupt	Active	Bankrupt
Return						
Active	1.00	0.73	0.37	0.75	0.03	−0.06
Bankrupt		1.00	0.33	0.53	0.17	−0.12
Volatility						
Active			1.00	0.61	0.24	−0.19
Bankrupt				1.00	0.25	0.25
Beta						
Active					1.00	0.16
Bankrupt						1.00
Market performance (composite indices)						
NYSE	0.48	0.19	−0.12	0.37	0.20	0.59
NASDAQ	0.49	0.13	0.00	0.33	−0.02	0.46

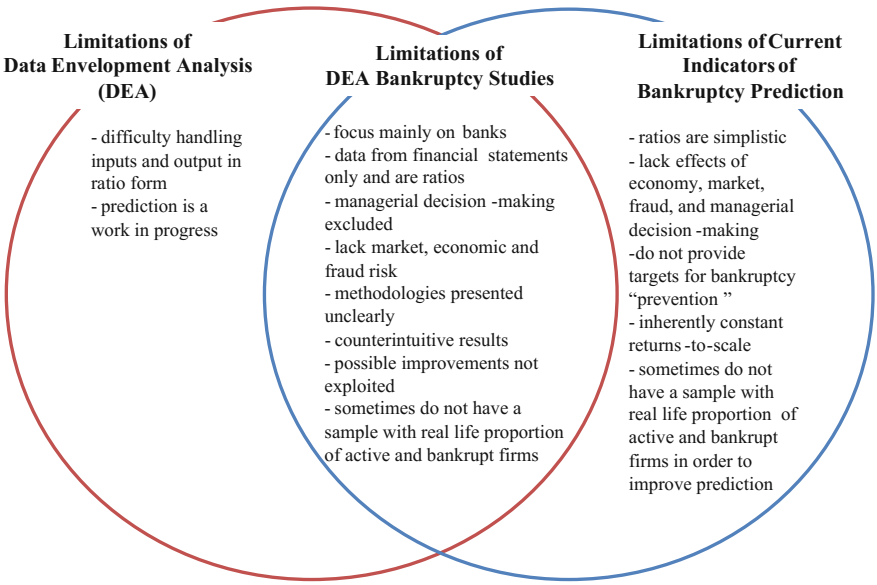


Fig. 13.1 Current limitations of DEA in bankruptcy prediction

of the SBM model was used for the analyses detailed in this paper. As such, any negative values had to be removed from the data. This was accomplished by defining new variables to house the negative values on the opposite side of the input/output designations. For example, if a DMU had an EBIT (earnings before interest and taxes - an output variable) of −$2M, this was replaced by zero and the DMU was given a value of +$2M for the created input variable, Negative EBIT.

Table 13.6 Confusion matrix

	Actual	
	+	−
Predicted		
+	True positive (TP)	False positive (FP) type I error
−	False negative (FN) type II error	True negative (TN)

The bankruptcy models considered were evaluated based on the type I and II error rates. Type I and II errors are the misclassification of bankrupt and active DMUs, respectively. In the context of this work, "*positive*" and "*negative*" results denote classifying a DMU as active and bankrupt, respectively. Thus a type I error would be a false positive, and a type II error a false negative. The error and overall success rates are defined as:

type I error rate = FP/(FP + TN),
type II error rate = FN/(FN + TP), and
success rate = (TP + TN)/(FP + FN + TP + TN),

where FP, FN, TP and TN are the number of false positive, false negative, true positive and true negative results, respectively. The various outcomes are summarized in the confusion matrix shown in Table 13.6.

The developed models are compared to the results of the public, non-manufacturing form of the Altman Z-score on the retail-apparel data set. The Altman model was chosen as the benchmark as it is a widely used standard in bankruptcy prediction, and this form of the model was the best performer of the three formulations on the data. As it is assumed that type I errors would be the more costly type, most of the focus is on comparing the type I error rates of the DEA models with those obtained from the Z-score. The type I error rates of the three Z-scores, as well as those from Altman's original paper (Altman 1968) are shown in Fig. 13.2.

DEA models generally classify DMUs by placing the obtained efficiency scores into zones by comparison with cut-off values. The focus of this work is a two-stage DEA model. The first-stage consists of various DEA models meant to reflect different aspects of the financial, MDM, and economic and market characteristics of the firm. The scores from these models were employed in a second-stage DEA model, and the DMUs were classified based on a layering approach. Layering is a DEA technique wherein the efficient DMUs are removed and the sample rerun, and is often helpful in discriminating between efficiency scores that are close in value, c.f. (Divine 1986; Thanassoulis 1999; and Paradi et al. 2004). Most studies only remove or "peel" off 2 or 3 layers of efficient DMUs. The novel approach employed here is to continue to remove layers until the remaining number of DMUs is less than minimum given by the accepted approach, i.e. $n \geq \max\{m \times s, 3(m + s)\}$, and to utilize the efficient layer number on which the DMU appears as the means to rank them and thus classify them as active or bankrupt. This approach is illustrated in Fig. 13.3.

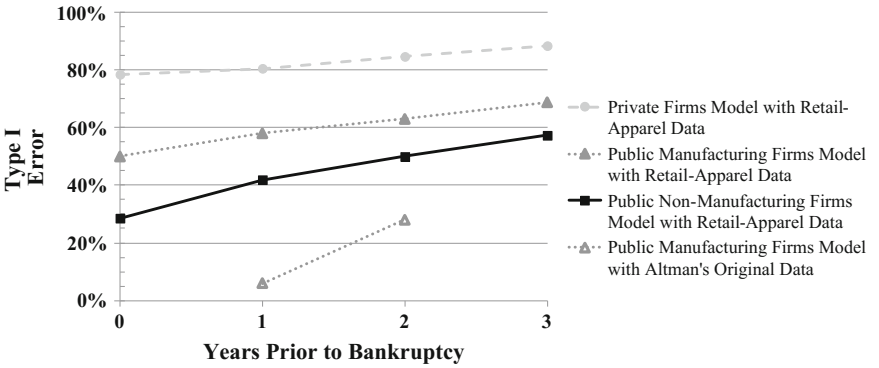


Fig. 13.2 Type I error from Z-Score by years prior to bankruptcy

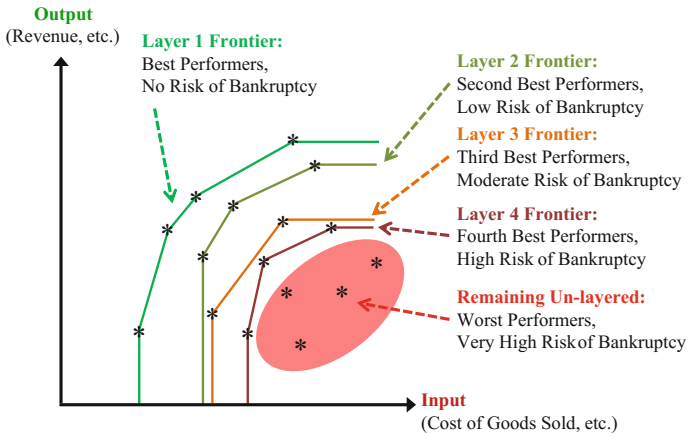


Fig. 13.3 Layering technique for DEA

13.5 First-Stage DEA Models

DEA can act as a multicriteria sorting tool with the objective function, i.e. relative efficiency score, serving as a benchmark to organize and classify firms by their level of health. Higher scores will generally be interpreted as a lower relative bankruptcy risk based on the given set of criteria. Thus, the selection of variables and their designation as inputs or outputs is of the utmost importance as this directly affects the determined frontier (i.e. benchmarks) from which scores are determined.

While most studies focus on measuring profitability and/or insolvency, a different approach was taken in this work: creating individual metrics that reflect a particular aspect of an annual report. In addition to considering the financial aspects of the firms, potential models reflecting management decision making and the

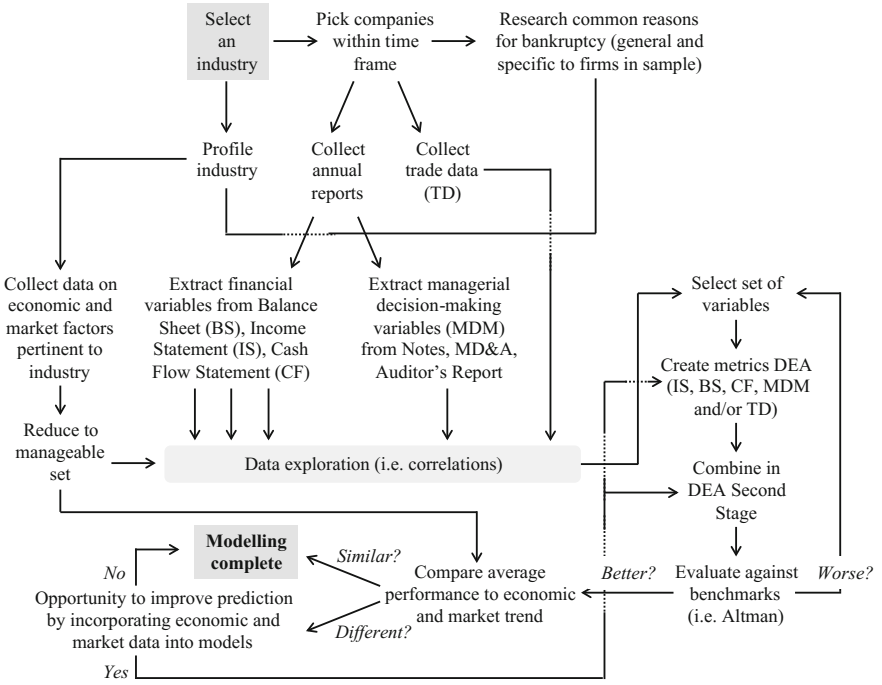


Fig. 13.4 Summary of methodology as applicable to all industries

economic and market conditions were also considered. While these aspects have not been typically considered in prior works, it was felt that they may capture additional important dimensions of the firms' health, and thus improve the predictions made. Figure 13.4 gives the flowchart of the overall methodology, as it could be applied to any industry.

Financial information is contained in the balance sheet, income statement and cash flow statement. For the retail-apparel industry, it was found that models generated from cash flow statement variables did not provide much discriminatory power, as score distribution produced was heavily bimodal, with over 90% of the scores less than 0.1. Three financial models were used: one for the income statement (IS) and one each representing assets (BSA) and liabilities (BSL) from the balance sheet. Many different combinations of variables were attempted, and several other models performed similarly to those chosen. Table 13.7 gives the inputs and outputs of each chosen model.

The MDM model used the six variables described in the data collection section as inputs (see Table 13.3), with a unit dummy output for all DMUs. A model based on the firms' trading performances using their returns, betas and volatilities produced similar score distributions as the cash flow statement model, i.e. heavily bimodal with little discriminatory power, and was not considered further.

Table 13.7 Inputs and outputs variables of IS, BSA and BSL (financial) models

Model	Inputs	Outputs
IS (income statement)	COGS (costs of goods sold)	NI (net income)
	ITE (income tax expense)	Revenue
	IE (net interest expense)	
	SGA (selling, general and administrative expenses)	
BSA (balance sheet—assets)	AR (accounts receivable)	Cash
	CA (current assets)	RE (retained earnings)
	Goodwill	SE (shareholders' equity)
	Inventory	
	MS (marketable securities)	
	PPE (net property, plant and equipment)	
	LTIS (long-term investment in securities)	
	TA (total assets)	
BSL (balance sheet—liabilities)	AP (accounts payable)	RE
	CL (current liabilities)	SE
	LTD (long-term debt)	
	CM (current maturities of LTD)	
	NPSTD (notes payable and short-term debt)	
	TL (total liabilities)	

Twenty overall market and economic factors had been selected, with data covering the period from 1988 to 2009. As each DMU is a year for this data, there were insufficient data points to run them together. Instead, they were divided into four DEA models representing separate indicators for: the general economy, the apparel industry, the housing market, and prices. A fifth indicator representing the general equity market performance was determined as the 45–55 weighted average of the normalized values of the NASDAQ and NYSE composite indices, where the weights were chosen based on the proportion of the firms studied listed on each exchange (OTC firms were grouped with the NYSE firms). The precise input and output variables for each of these indicator DEA models are presented in Table 13.8. The five scores were averaged to provide an overall market and economic indicator.

All of the first-stage variable sets for the three financial models (IS, BSA, and BSL) were run in regular and inverse DEA models—an inverse DEA model has the inputs and outputs from a regular DEA model reversed, i.e. an input variable in the original model is an output in the inverse model. Based on the DEA results, a single set of cut-off values was chosen that was found to work well, in sample, for all the financial models. For the regular models, scores below 0.4 were deemed bankrupt, those above 0.7 were deemed active, and DMUs with intermediate scores were not classified by the model. A single cut-off value of 0.7 was used for the inverse DEA

Table 13.8 Description of market and economic (ME) factors models

Indicator model	Inputs	Outputs
General economic	Unemployment rate	GDP growth rate
	Debt as % of GDP	Inflation
		Prime interest rate
Apparel industry	Dummy (DMU input = 1)	Personal consumption expenditures: clothing
		GDP: clothing
		Apparel labour productivity
		Apparel imports
Housing market	Median months for a sale	Housing unit starts
Prices	Oil prices	Dummy (DMU output = 1)
	Cotton prices	
General market performance	Score is 45–55 weighted average of normalized NASDAQ and NYSE composites	

models run, and DMUs with scores below 0.7 were considered active while those above 0.7 bankrupt.

In general, the distributions of scores obtained from the inverse models run were heavily bimodal, with the majority (90%+) of scores below 0.1, some efficient DMUs, and very little probability mass in between. This distribution led to poor discriminatory power and very high type I error rates, which is considered particularly undesirable as type I errors are generally deemed more expensive than type II errors in bankruptcy prediction. The normal DEA models had lower type I errors and higher overall success rates, but also resulted in some DMUs being unclassified.

The scores from the regular DEA models were only weakly correlated with those of the corresponding inverse models. As such, combining the regular and inverse models both in series (using the inverse model on the unclassified DMUs from the normal model) and in parallel (requiring agreement between the two models) were also investigated. The performance of these combinations in terms of error and classification rates was intermediate between that of the individual models, without any particular improvement to overall performance. Thus, inverse DEA models were not considered further.

The benchmark Altman Z-score had type I error rates of 30, 42, 50 and 60%, zero to three years prior to bankruptcy respectively. The type II error rate was relatively consistent at 12% across the same period. The results from the individual financial statement DEA models were used to classify the DMUs by two means. The scores were compared to fixed classification cut-offs (0.4 and 0.7). The overall classification results, covering all predictions for the year of bankruptcy up to three years prior, are summarized in Table 13.9.

With the exception of the MDM model, the DEA models had lower overall type I errors but higher type II errors than the Z-score. The MDM had extremely high

Table 13.9 Summary of first-stage results for IS, BSA, BSL and MDM models

Model	Rates (%)			
	Type I error	Type II error	Success	Unclassified
IS	11.9	31.8	69.7	21.3
BSA	6.7	77.7	27.7	21.0
BSL	0	87.3	19.9	22.8
MDM	82.6	2.8	92.9	15.7

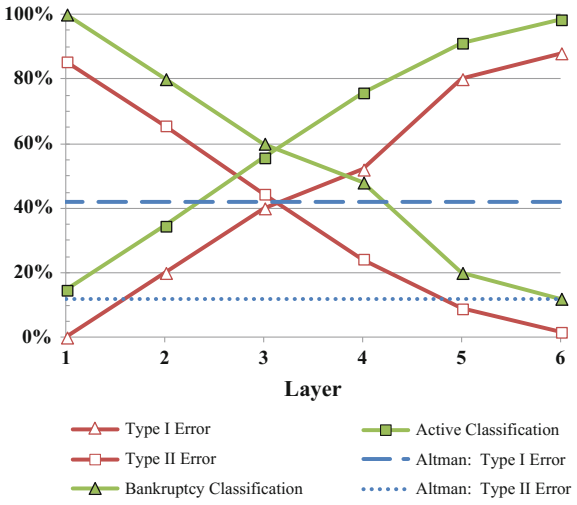


Fig. 13.5 Prediction by IS model from one year back (A DMU was classified bankrupt up to a year prior to filing)

type I errors, but low type II errors. The unclassified rates were approximately 20% for all the models.

The process of peeling back layers until insufficient DMUs remained was also applied to each model. Figure 13.5 shows the variation of error rates with cut-off layer for the IS model, one year prior to bankruptcy, where a DMU is considered bankrupt up to one year prior to filing. All DMUs on the efficient frontiers from the first to the cut-off peel are considered active, and all others are classified as bankrupt. If only the first frontier is used, almost all the efficient DMUs will be active, resulting in a very low type I error rate. However, most DMUs, including many of the active ones, will not be on the first frontier, causing a very high type II error rate. As the number of layers used for classification increases, the type I error rate increases, and type II error rate decreases.

Although type I errors are probably more costly than type II errors, the point where type I and II error rates intersected was chosen as the cut-off layer since no explicit information about their relative costs were available. For example, for the IS model this cut-off would be layer 3 (Fig. 13.5). The 1-year prior to bankruptcy results for the 4 models are given in Table 13.10. The IS model had comparable type I errors and worse type II errors than the Z-score (c.f. type I and II error rates of

Table 13.10 Cutoff layer and type I error, type II error and success rates for first-stage models

Model	Cutoff layer	Rates (%)		
		Type I error	Type II error	Success
IS	3	40	43	58
BSA	4	18	24	77
BSL	8	22	23	77
MDM	2	34	35	65

Table 13.11 Correlation between first-stage DEA scores (all correlations were significant at 0.01 level)

	IS	BSA	BSL	MDM
Income statement	1	0.49	0.47	0.14
Balance sheet assets		1	0.71	0.08
Balance sheet liabilities			1	0.15
Managerial decision-making				1

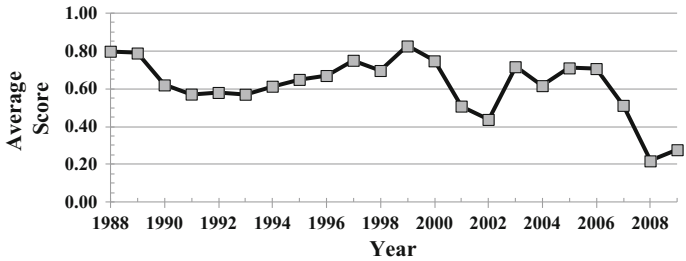


Fig. 13.6 All-encompassing market and economic indicator by year

42 and 12% respectively). The error rates for the other three models fell in between those of Z-score, i.e. lower type I errors, higher type II errors. Note that in some cases, better performance relative to the Z-score could have been achieved with different cut-off layers. For example, if layer 7 was employed as the cut-off for the BSL model, it would have the same type II error rate, but lower type I errors than the Altman model. Although not shown, the error rates for predictions two and three years prior to bankruptcy were somewhat higher than those one year prior, as expected.

The correlations between the efficiency scores from the four models are given in Table 13.11. The correlation coefficients were all significant at the 1% level, and were generally weak. The BSA and BSL models had a correlation of 0.71, and all other pairs were below 0.5. This supports the hypothesis that the models represent different aspects of company health, and could be considered together to provide a more complete view.

Figure 13.6 gives the determined overall market and economic indicator. Note that the local minimums roughly correspond to the three US recessions that occurred during this period: the savings and loans crisis (1990–91), the tech bubble collapse and Sept. 11 attacks (2001–02), and the subprime mortgage crisis (2007–09).

13.6 Second-Stage DEA Models

Three second-stage DEA models were investigated. These were DEA (VRS, output-oriented SBM) models with a dummy unit input, and first stage scores as outputs. The first model only considered the scores from the financial models (i.e. IS, BSA, and BSL). The second model used the financial and MDM model scores as outputs, and the final model used all of the financial and MDM model scores and the overall market and economic (ME) indicator as outputs (all DMUs for a particular year have the same value for the ME indicator).

Zone classification results of the second-stage DEA models are summarized in Table 13.12. They showed that adding MDM scores to the financial scores generally improved results, reducing type I errors from 32 to 21%, with slight effects on type II errors (increased from 16 to 19%), success (decreased from 83 to 81%) and unclassified rates (decreased from 33 to 32%). Adding the overall ME indicator to the model proved detrimental, increasing the type I error rate to 98%. These results partially support one of the main hypotheses of this paper, that considering the MDM information improves model predictions, whereas adding the ME indicator did not. As such, further layering classification proceeded with the second model, i.e. incorporating financial and MDM first stage DEA scores.

Classification using layering was chosen over that by zones for the final model because it had better discrimination, resulted in no unclassified DMUs, and required a less subjective choice of cut-offs points (choosing the cut-off layer to equate the two error rates versus choosing high and low score zones). Additionally, a novel technique was developed to translate the layer classifications into an efficiency score-type measure. This layered score is defined as Eq. 13.3:

$$0 < \text{Layered Score} = \frac{N + 1 - L}{N} \leq 1, \tag{13.3}$$

Table 13.12 Second-stage model predictions with classifications by zones

%	IS, BSA, BSL	IS, BSA, BSL, MDM	IS, BSA, BSL, MDM, ME
TP rate	84.0	80.7	100.0
FP rate	32.3	20.7	97.7
Type I error	32.3	20.7	97.7
Type II error	16.0	19.3	0
Success rate	82.9	80.6	93.6
Unclassified	33.1	31.5	4.0

Table 13.13 Correlation of first-stage layered scores

	IS	BSA	BSL	MDM
Income statement (IS)	1	0.23	0.14	0.06
Balance sheet assets (BSA)		1	0.74	0.12
Balance sheet liabilities (BSL)			1	0.19
Managerial decision-making (MDM)				1

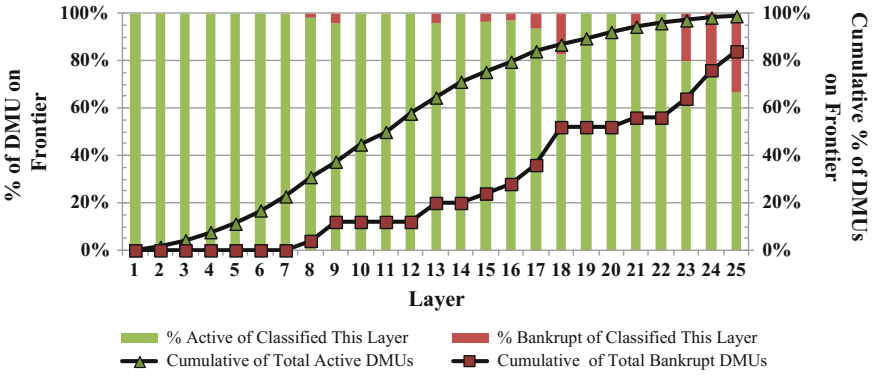


Fig. 13.7 Classification by layer in percentages, one year back

where N is the total number of layers (i.e. until the remaining number of DMUs is less than the accepted minimum to run another analysis), and L is the layer on which the DMU appears on the frontier (L = N + 1, i.e. layered score = 0, for those DMUs that were still inefficient after layer N). This score definition had the desired property of varying from a minimum of zero for those DMUs not on any layer to a maximum of one for those on the first layer, and hence lower values can be expected to correspond to increased risk of bankruptcy. Layered scores for the four first-stage DEA models (i.e. financial statement and MDM models) were generally weakly correlated (max. of 0.74 between BSA and BSL, less than 0.25 otherwise), and are given in Table 13.13.

The cumulative total % curves of active and bankrupt DMUs identified up to a given layer (Fig. 13.7) show that both increase as the number of layers increases, as expected. Also, as expected, as the layer number increases, the active-to-bankrupt ratio of DMUs on that frontier tends to decrease.

The error rates of the layering predictions are given in Table 13.14. Overall, the second-stage DEA model provided comparable error rates to the BSA and BSL first-stage models (a little better 1–2 years before bankruptcy, a little worse 3 years before), and lower errors than the IS and MDM models alone. The second-stage model's type I error rates are lower than those from the Altman Z-score, for each of 1–3 years before bankruptcy, while its type II error rates are higher (choosing the intersection of the two errors rates as the cut-off number of layers). As type I errors

Table 13.14 Accuracy of classification by layering of second stage model and individual metrics

	One year back			Two years back			Three years back		
	Accuracy (%)	Error (%)	Layer	Accuracy (%)	Error (%)	Layer	Accuracy (%)	Error (%)	Layer
Second stage	80	20	16	73	27	14	69	31	13, 14
IS	58	42	3	55	45	3	55	45	3
BSA	78	22	4, 5	72	28	4	70	30	3, 4
BSL	78	22	4	72	28	7, 8	70	30	7
MDM	65	35	2	60	40	2	61	39	2

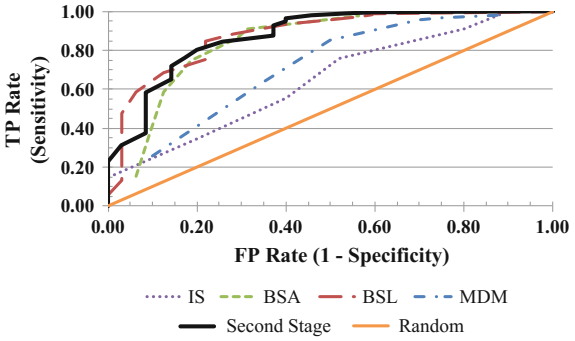


Fig. 13.8 ROC curves for one year back results

are assumed to be more costly, the DEA model may still outperform the Z-score even if its error rate lies between the Altman type I and II error rates. Furthermore, unlike the Z-score, DEA layering classification results in no unclassified DMUs.

The various DEA models are further compared by examining the receiver operating character (ROC) curves, which plot the TP rate (=1 − type II error rate) versus the FP rate (=type I error rate), in Fig. 13.8. The ideal for these curves would be the upper left corner (i.e. no errors), and employing only a few layers with any of the models will tend to result in a location near the origin. By these curves (both visual inspection and computing areas under the curves), the 2-stage and BSL models are comparable and the best performers; the BSA is a slightly worse performer by this measure, mainly in the initial (left side) of the curve, which is the more important section. All DEA models performed better than random guessing (the y = x line). Also if the intersection type I and II errors is chosen as optimal number of layers, the intersection of ROC plot with y = 1 − x curve would give this.

Figure 13.9 compares the zone and layering classification predictions, one year prior to bankruptcy, of the second-stage DEA model. It had a total of 25 layers, and layer 16 was the cut-off. The correlation between the two sets of scores was 0.82, and was stronger for the higher scores, with more variation in the number of peels for the lower raw scores. Interestingly, the data is bounded from below by the

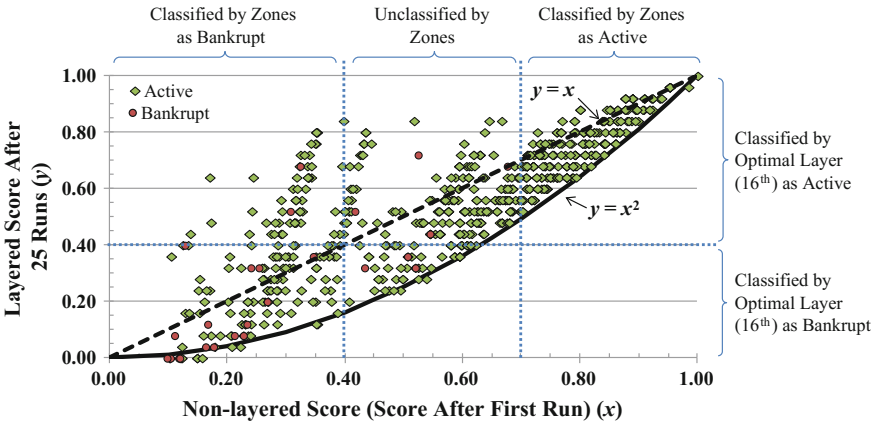


Fig. 13.9 Correlation between layering and non-layering techniques (one year prior to bankruptcy)

Table 13.15 Performance of layering and non-layering techniques

Evaluation (%)	Layering	Non-layering: two zones	Non-layering: cut-off of 0.8	Non-layering: cut-off of 0.7	Non-layering: cut-off of 0.6
TP rate	80.3	80.7	37.7	54.6	66.9
FP rate	20.0	20.7	0	4.0	20.0
Type I error	20.0	20.7	0	4.0	20.0
Type II error	19.7	19.3	62.3	45.4	33.1
Success rate	80.3	80.6	39.9	56.1	67.3
Unclassified	0	31.5	0	0	0

$y = x^2$ curve. The error rates from layering are comparable to those from using two zones, although it was more effective as it classified all DMUs. These results, as well as those using various cut-offs for a single zone are given in Table 13.15.

13.7 Discussion

Like most prior bankruptcy studies, the layering DEA results provide a yes/no classification, but not a probability of bankruptcy. However, the scores can be translated into probabilities. Figure 13.10 shows that as the layered scores decrease (i.e. higher layer number), the number of bankrupts on that layer increase, as expected. The actual probabilities of active and bankrupt DMUs on each layer (from 26, i.e. not on any frontier, to 1) are tabulated in Table 13.16, and used to

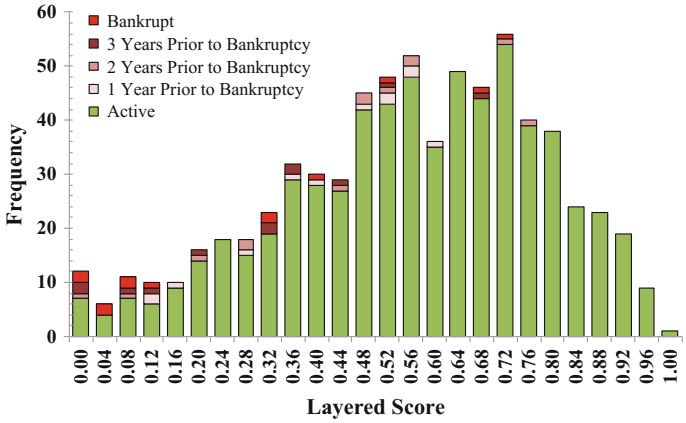


Fig. 13.10 Distribution of second-stage layered scores

generate CDFs for the bankruptcy probabilities. A second order polynomial was fitted, i.e.

$$y = 0.47x^2 - 0.78x + 0.39$$

which matched the actual data well and is shown in Fig. 13.11.

Window analysis was used for validation of the bankruptcy predictions. The 2-stage DEA model was used in six windows, from 1996 to each of 2004–2009, to fit a bankruptcy probability polynomial. These were then used to predict bankruptcies in the next two years after the window in five of the six cases. As the fitted CDF and the underlying empirical data were not monotonically decreasing for the shortest (1996–2004) window, the fitted function could not be reliably used for bankruptcy prediction. In-sample classification results and bankruptcy probability functions for each of the windows are summarized in Table 13.17. Accuracy decreased as the number of years (i.e. DMUs) decreased, as DEA generally performs better with more DMUs, but was consistent (75–80%), indicating the selection of variables and metrics used was robust. With the exception of the last window, the second-order CDFs fitted for each window had good fits, with results similar to that shown in Fig. 13.11.

The out-of-sample predictions were carried out by converting the scores of all the remaining DMUs in the last year of the window into bankruptcy probabilities (y) and noting whether they declared bankruptcy in either of the next two years. In general, the forward predictions were less reliable as the model timeframe shortened, (i.e. lower accuracy as the number of DMUs in the fitting window decreased).

For the window ending in 2009, there were 39 firms in the final year, all of which were active. Twenty-three had scores above 0.6 ($y < 3\%$), all of which remained active. Of the 3 that had scores below 0.3 ($y > 13\%$), one went bankrupt, one stayed active, and one underwent a restructuring. Although they did not declare

Table 13.16 Probabilities of bankruptcy (B) and non-bankruptcy (NB)

Layer	Score	Actual								Estimate	
		Count of DMUs on frontier			Cumulative count of DMUs on frontier			P(NB) (%)	P(B) (%)	Second order polynomial function	
		Total	NB	B	Total	NB	B			P(NB) (%)	P(B) (%)
26	0.00	–	–	–	–	–	–	–	–	60.8	39.2
25	0.04	6	4	2	6	4	2	66.7	33.3	63.9	36.1
24	0.08	11	7	4	17	11	6	64.7	35.3	66.8	33.2
23	0.12	10	6	4	27	17	10	63.0	37.0	69.5	30.5
22	0.16	10	9	1	37	26	11	70.3	29.7	72.1	27.9
21	0.20	16	14	2	53	40	13	75.5	24.5	74.6	25.4
20	0.24	18	18	0	71	58	13	81.7	18.3	76.9	23.1
19	0.28	18	15	3	89	73	16	82.0	18.0	79.0	21.0
18	0.32	23	19	4	112	92	20	82.1	17.9	81.0	19.0
17	0.36	32	29	3	144	121	23	84.0	16.0	82.9	17.1
16	0.40	30	28	2	174	149	25	85.6	14.4	84.6	15.4
15	0.44	29	27	2	203	176	27	86.7	13.3	86.1	13.9
14	0.48	45	42	3	248	218	30	87.9	12.1	87.5	12.5
13	0.52	48	43	5	296	261	35	88.2	11.8	88.8	11.2
12	0.56	52	48	4	348	309	39	88.8	11.2	89.9	10.1
11	0.60	36	35	1	384	344	40	89.6	10.4	90.8	9.2
10	0.64	49	49	0	433	393	40	90.8	9.2	91.7	8.3
9	0.68	46	44	2	479	437	42	91.2	8.8	92.3	7.7
8	0.72	56	54	2	535	491	44	91.8	8.2	92.8	7.2
7	0.76	39	39	0	574	530	44	92.3	7.7	93.2	6.8
6	0.80	38	38	0	612	568	44	92.8	7.2	93.4	6.6
5	0.84	24	24	0	636	592	44	93.1	6.9	93.5	6.5
4	0.88	23	23	0	659	615	44	93.3	6.7	93.4	6.6
3	0.92	19	19	0	678	634	44	93.5	6.5	93.1	6.9
2	0.96	9	9	0	687	643	44	93.6	6.4	92.7	7.3
1	1.00	1	1	0	688	644	44	93.6	6.4	92.2	7.8

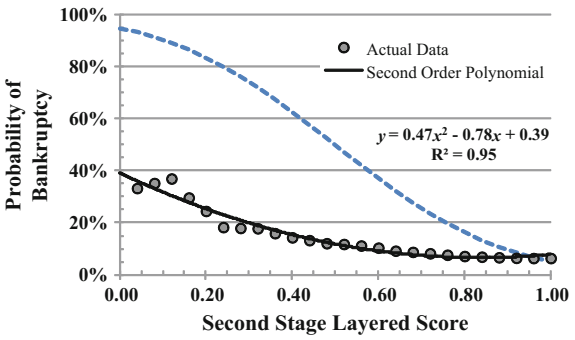


Fig. 13.11 Probability of bankruptcy

Table 13.17 Classification by layering and fitted second order bankruptcy probability polynomials from one year back for different windows

Time window	Total number of DMUs (active, bankrupt)	Total number of peels	Accuracy (%)	Error (%)	Layer	Probability of bankruptcy (y) from layered score (x)
1996–2009	701 (651, 50)	26	80	20	16	$y = 0.54x^2 - 0.81x + 0.32$
1996–2008	584 (543, 41)	23	76	24	14	$y = 0.63x^2 - 0.91x + 0.34$
1996–2007	532 (492, 40)	22	76	24	13, 14	$y = 0.24x^2 - 0.41x + 0.20$
1996–2006	476 (439, 37)	21	76	24	13	$y = 0.16x^2 - 0.31x + 0.17$
1996–2005	409 (379, 30)	19	75	25	11, 12	$y = 0.05x^2 - 0.12x + 0.10$
1996–2004	357 (331, 26)	18	77	23	11, 12	N/A

bankruptcy, most of the 13 firms with intermediate layered scores underwent transitions (e.g. reached new credit agreements, closed stores, merged and/or went private), indicating that they were less stable and higher risk than those with higher scores.

There were 52 firms in 2008, 1 of which was bankrupt. Twenty-nine active firms had scores above 0.6 ($y < 2\%$), of which 27 stayed active and 2 went bankrupt at the onset of the recession. Seven active firms and the bankrupt firm had scores below 0.4 ($y > 8\%$). Of the active firms, three stayed active, one went private, one changed its name while two went bankrupt. The firm that was already bankrupt in 2008 emerged from bankruptcy during the two year period. Of the 15 firms with intermediate scores, 10 remained active and 5 experienced transitions (new credit agreements or store closings).

There were 55 active and 1 bankrupt firm in 2007. All but 1 of the 35 firms with scores above 0.6 ($y < 4\%$) remained active over the two years. Five active firms has scores below 0.3 ($y > 10\%$); 3 of these went bankrupt, 1 went private and 1 stayed active. The bankrupt firm also had a score below 0.3, and emerged from bankruptcy. Ten of the 15 firms with intermediate scores remained active, and 2 went bankrupt.

All 50 firms in 2006 were active. The thirty firms that had scores above 0.6 ($y < 4\%$) remained active over the two years. Of the 6 firms with scores below 0.3 ($y > 9\%$), 2 went bankrupt and 4 stayed active. One of the 14 firms with intermediate scores went bankrupt and the rest remained active. There were 47 active firms in 2005; of these all 26 that had scores above 0.6 ($y < 5\%$) remained active. Five firms had layered scores below 0.3 ($y > 7\%$), and 16 had scores in between. One firm from each of those groups went bankrupt, and the remainder stayed active.

Another aspect of the DEA model in comparison to other bankruptcy prediction methodologies is that the DEA model can also be used to provide comparison and improvement targets for the DMUs. From the results of the first run (layer) of the second-stage model, the required improvements for the individual scores to attain overall efficiency can be ascertained. The first-stage model results for those scores give the necessary improvement targets for the individual inputs and outputs. By concentrating on the company aspects, represented by first-stage scores, and the individual inputs and outputs within those models, allowing the greatest potential improvements, areas for these companies to focus upon in order to improve are identified. This procedure can also be applied to active DMUs as a means to improve profitability.

Overall, the bankrupt DMUs generally required improvement in all four first-stage metrics to achieve efficiency. For the first-stage IS model, the main improvement identified for most bankrupt DMUs was to decrease net interest and income tax expenses, and increase net income. In the BSA model, bankrupt DMUs should look to reduce accounts receivable, inventories, PPE and total assets, and increase cash, retained earnings and shareholders' equity to become efficient. Improving all variables was found to be important in the BSL model, and reducing the management turnover, legal proceedings, related party transactions, and pension variables were identified as important in the MDM model.

Including the ME factor made predictions worse. One hypothesis to explain this is that the information in the ME data is already being captured by the other financial and MDM data included. The plots of the ME overall indicator and the average DEA layered scores (Fig. 13.12) are similar. The various tests performed to test the null hypothesis that there is no significant difference between them are summarized in Table 13.18. All of these showed no significance, except that their correlation was weak, and also their annual direction of change (up or down), which is probably more important than the actual magnitude of the scores, was the same in 10 out 12 years. Thus it was concluded that the hypothesis was true.

In the data sample studied, the ratios of companies that were active (never bankrupt) to those that were bankrupt at least once was 61–18 (3.4–1), and active to bankrupt DMUs was 651–50 (13–1). Figure 13.13 demonstrates that by reducing this ratio of companies from 3.4–1 to 1–1, classification improved, as was previously seen in other bankruptcy studies. In fact at ratios below 2–1, perfect

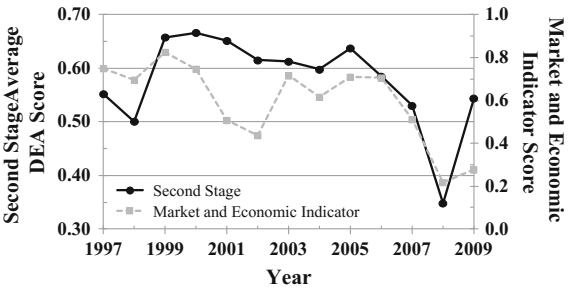


Fig. 13.12 Average scores/indicator values by year

Table 13.18 Tests of similarity between second-stage layered scores and overall ME indicator

Evaluation		Conclusion
Significance in difference of means		
Mann-Whitney U test	0.29	No significant difference between means of scores
T-test	0.77	
ANOVA	0.77	
Correlation		
Pearson	0.62	Weak correlation between scores
Spearman	0.50	
Difference in area		
Area under the Curves	3.6%	No significant difference
Coefficient of Divergence (Wongphatarakul et al. 1998)		
Coefficient of divergence	0.15	Similarity

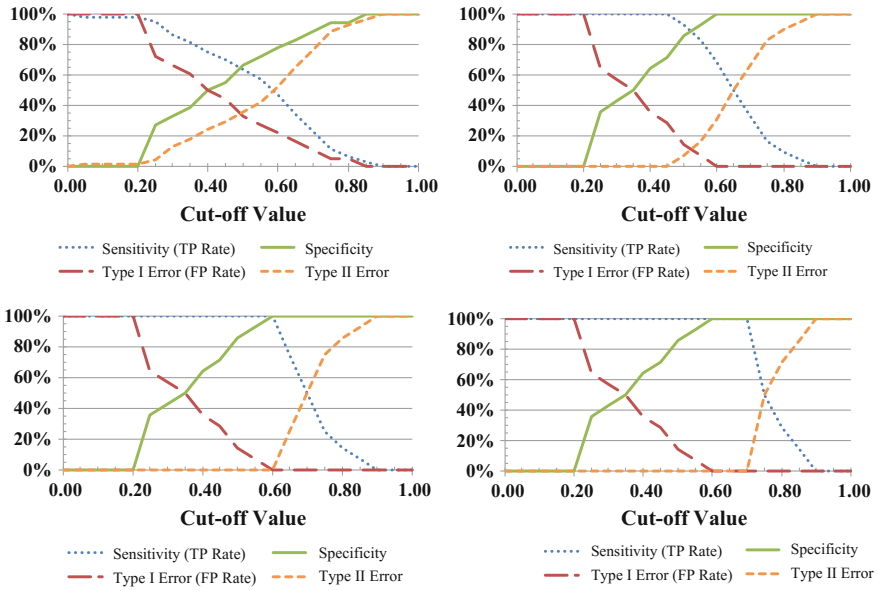


Fig. 13.13 Classification with active-to-bankrupt firms ratios of 3.4–1, 3–1, 2–1 and 1–1 respectively

separation of the active and bankrupt firms was achieved using the classification by zones depicted. It should be noted that as the firms were being compared, it is their average score over all years for which they had data that was plotted and used for classification. Also, when the active-to-bankrupt ratio was reduced, the active firms used were not random, but instead the firms were those with the highest average scores—it would be assumed that this improves the results obtained.

13.8 Conclusions and Future Work

This paper studied bankruptcy prediction in 85 US retail-apparel firms that were traded for some continuous period from 1996–2009. Financial, management decision making (MDM), and key market and economic data were considered as possible explanatory variables. They were found to be largely uncorrelated with each other, but individually poor at predicting bankruptcy. This led to the main hypothesis that including variables for MDM and market and economic factors would improve prediction compared to models that considered financial data alone; for the retail-apparel industry, only inclusion of the former was found to be beneficial; however, market and economic data could be helpful in models for other industries.

The developed methodology involved creating metrics (DEA SBM models) based on different aspects of the annual report, and combining them into a second-stage SBM model. An extended DEA layering technique was used to classify the DMUs, and the results compared favorably to the public, non-manufacturing form of the Altman Z-score used as a benchmark. The final model had a type I error rate of 20% one year prior to bankruptcy, compared to 42% for the Z-score. A layered efficiency based on the frontier number where a DMU appeared was developed, and it was found that a 2nd order polynomial of this score could be used to provide bankruptcy probabilities in addition to a simple bankrupt/non-bankrupt classification. Finally, as the model is DEA-based, the efficient projection can provide explicit improvement targets for individual data items for both bankrupt and active firms.

The developed methodology could easily be adapted to other industries. Although the chosen variables for the individual metrics, and even the metric themselves (e.g. including market and economic factors, or a metric from the cash flow statement) would likely be different, the selection process was based on reducing an all-inclusive list to those which are shown to have an effect, based on the tests of correlation and similarity. Thus, it was not overly subjective.

Other extensions to this work could involve incorporating corporate governance risk indicators and other data not captured by the financial statements, or trends or some other means to reflect or calibrate for time, since the data from companies do not overlap entirely over the same time period. The differing costs of type I and II errors could be determined and accounted for in determining the model cut-off layers, or the percentage losses in bankruptcies could be considered as the optimizing objective instead of the correct classification of bankrupt and non-bankrupt firms. As there was some variation in the zone and layering scores for the lower scores, there might be some means to combine the two to improve predictions.

Furthermore, credit ratings determined by agencies such as S&P and Moody's could be considered in model extensions, especially given the importance placed by investors on their work with regards to default risk. These ratings could be incorporated as model variables, or the default probabilities determined from the layering results could be compared to historical default probabilities implied by the firm's

credit ratings, either across all industries, or preferably those specific to the industry under study.

References

US Courts, Bankruptcy Statistics (2010). http://www.uscourts.gov/bnkrpctystats/bankruptcystats.htm

Bradley DB, Cowdery C (2004) Small business: causes of bankruptcy. Internal document of UCA Small Business Advancement National Center, pp 205–219

Ravi Kumar P, Ravi V (2007) Bankruptcy prediction in banks and firms via statistical and intelligent techniques—a review. Eur J Oper Res 180:1–28

Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J Financ 23(4):589–609

Hanson RO (2003) A study of Altman's revised four-variable Z″-score bankruptcy prediction model as it applies to the service industry (Edward I. Altman). Dissertation abstracts international, pp 63–12A, 4375

Ohlson JA (1980) Financial ratios and the probabilistic prediction of bankruptcy. J Acc Res 18 (1):109–131

Beaver W (1966) Financial ratios as predictors of failure. Empirical Res Account Sel Stud Suppl J Account Res 4:71–111

Wilcox JW (1971) A simple theory of financial ratios as predictors of failure. J Account Res 9 (2):389–395

Deakin EB (1972) A discriminant analysis of predictors of business failure. J Account Res 167–179

Edmister RO (1972) An empirical test of financial ratio analysis for small business failure prediction. J Financ Quant Anal 7(2):1477–1493

Blum M (1974) Failing company discriminant analysis. J Account Res 1–25

Libby R (1975) Accounting ratios and the prediction of failure: some behavioural evidence. J Account Res 150–161

Moyer R (1977) Forecasting financial failure: a re-examination. Financ Manage 6(1):11

Cooper WW, Seiford LM, Tone K (2007) Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software. Springer, New York

Tone K (2001) A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res 130(3):498–509

Banker RD, Charnes A, Cooper WW, Swarts J, Thomas DA (1989) An introduction to data envelopment analysis with some of its models and their uses. Res Govt Nonprofit Acc 5:125–163

Hollingsworth B, Smith PC (2003) The use of ratios in data envelopment analysis. Appl Econ Lett 10:733–735

Emrouznejad A, Amin GR (2007) DEA models for ratio data: Convexity consideration. Appl Math Modell 33(1):486–498

Barr RS, Seiford LM, Siems TF (1993) An envelopment-analysis approach to measuring the managerial efficiency of banks. Ann Oper Res 45:1–19

Barr RS, Seiford LM, Siems TF (1994) Forecasting bank failure: a non-parametric frontier estimation approach. Recherches Economiques de Louvain 60(4):417–429

Kao C, Liu ST (2004) Predicting bank performance with financial forecasts: a case of Taiwan commercial banks. J Bank Finance 28:2353–2368

Cielen A, Peeters L, Vanhoof K (2004) Bankruptcy prediction using a data envelopment analysis. Eur J Oper Res 154(2):526–532

Premachandra M, Chen Y, Watson J (2011) DEA as a tool for predicting corporate failure and success: a case of bankruptcy assessment. Omega 39(6):620–626

Xu X, Wang Y (2009) Financial failure prediction using efficiency as a predictor. Expert Syst Appl 36:366–373

Paradi JC, Asmild M, Simak PC (2004) Using DEA and worst practice DEA in credit risk evaluation. J Prod Anal 21:153–165

Sueyoshi T (1998) DEA-discrimant analysis in the view of goal programming. Eur J Oper Res 22:349–376

Sueyoshi T, Goto M (2009) Methodological comparison between DEA (data envelopment analysis) and DEA-DA (discriminant analysis) from the perspective of bankruptcy assessment. Eur J Oper Res 199:561–575

Stickney CP, Brown P, Wahlen JM (2006) Financial reporting, financial statement analysis, and valuation: a strategic perspective. 6th edn. South-Western College Pub, Mason

Gill JO (1994) Financial basics of small business success. Crisp Publications, USA

Divine JD (1986) Efficiency analysis and management of not for profit and governmentally regulated organizations. Ph.D. dissertation. Graduate School of Business, University of Texas, Austin

Thanassoulis E (1999) Setting achievement targets for school children. Educ Econ 7(2):101–119

Wongphatarakul V, Friendlander SK, Pinto JP (1998) A comparative study of $PM_{2.5}$ ambient aerosol chemical databases. Environ Sci Technol 32:3926–3934

Chapter 14
Banking Crises, Early Warning Models, and Efficiency

Pavlos Almanidis and Robin C. Sickles

Abstract This paper proposes a general model that combines the Mixture Hazard Model with the Stochastic Frontier Model for the purposes of investigating the main determinants of the failures and performances of a panel of U.S. commercial banks during the financial crisis that began in 2007. The combined model provides measures of the probability and time to failure conditional on a bank's performance and vice versa. Both continuous-time and discrete-time specifications of the model are considered in the paper. The estimation is carried out via the expectation-maximization algorithm due to incomplete information regarding the identity of at-risk banks. In- and out-of-sample predictive accuracy of the proposed models is investigated in order to assess their potential to serve as early warning tools.

Keywords Financial distress · Panel data · Bank failures · Semiparametric mixture hazard model · Discrete-time mixture hazard model · Bank efficiency

JEL Classification Codes C33 · C41 · C51 · D24 · G01 · G21

14.1 Introduction

In light of the recent 2007–2011 financial meltdown in the United States (U.S.), during which more than 400 banks and thrifts failed, that is were forced into closure by regulatory agencies,[1] the need for effective regulations and intervention policy that would identify and resolve future crises and undertake prompt corrective

[1]According to the Federal Deposit Insurance Corporation's Failed Bank List available at: http://www.fdic.gov/bank/individual/failed/banklist.html.

P. Almanidis (✉)
International Transfer Pricing Services, Ernst & Young LLP, Toronto, Canada
e-mail: pavlos.almanidis@gmail.com

R.C. Sickles
Department of Economics, Rice University, Houston, USA

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_14

actions to resolve such crises with minimal cost in a timely fashion has been recognized as essential to the health of the U.S. economy. In this paper we only consider failed banks as ones that appear on the FDIC's failed bank list and ceased their operation due to reasons other than merger or voluntary liquidation, or that remained inactive or no longer were regulated by the Federal Reserve. The 2007–2011 financial crisis, which originated from the secondary market for residential mortgage-backed securities (RMBS) immediately after the collapse of the housing bubble in 2006, caused severe losses to banks and in particular large banks, which were highly involved in the RMBS market. At the same time and as a result of the large banks' widespread distress and contagion effects, the number of problem banks on the watch list maintained by the Federal Deposit Insurance Corporation (FDIC) dramatically increased. Systemically important financial institutions at risk, commonly described as too-big-to-fail, received heavy doses of government funds through the Troubled Asset Relief Program (TARP) from regulatory authorities who apparently believed that the banks' failures would impose greater systemic risk that could substantially damage the economy and lead to conditions similar to, or possibly exceeding, those of the Great Depression. The financial crisis footprint was not the same across the states. Those that experienced the most failures were California, Florida, Georgia and Illinois, accounting for more than half of the failures in the U.S.

Banking crises are not a new phenomena in the U.S. economy[2] and regulatory authorities have always considered banking failures as a major public policy concern, because of the special role that banks play in the economic network and in the implementation of an effective monetary policy. The distinguishing characteristic of the banking crisis of 2007–2011 from those in the 1980s and 1990s, however, is that failures were not limited to small financial institutions. Rapid credit expansion and low quality loans and investments made during a period of economic expansion mainly took their toll on large multi-billion dollar financial institutions. Approximately one in five banks that failed had assets of over $1 billion and in 2008 thirty-six percent were large banks, among them the largest bank failure in the history of the U.S., that of Washington Mutual with $307 billion in assets.[3] That same year saw Lehman Brothers file for Chap. 11 bankruptcy protection and IndyMac Bank with $32 billion in assets taken over by the FDIC.[4] These large financial institution failures created large uncertainties about the exposure of other financial institutions (healthy and troubled) to additional risks, reduced the availability of credit from investors to banks, drained the capital and money markets of

[2]The Great Depression of 1930s and savings and loan (S&L) crisis of the 1980s and 1990s are the two most obvious examples from the last century.

[3]Continental Illinois Bank and Trust Company of Chicago failed in 1984 and had one-seventh of Washington Mutual's assets.

[4]Chapter 11 permits reorganization under the bankruptcy laws of the United States. A financial institution filling for Chap. 11 bankruptcy protection usually proposes a plan of reorganization to keep its business alive and pay its creditors over time.

confidence and liquidity, triggered the failure of smaller community banks,[5] and raised fears of severe instability in the financial system and the global economy.

In the U.S., the FDIC and state banking regulatory authorities are responsible for the identification and resolution of insolvent institutions. A bank is considered at a risk of immediate closure if it is unable to fulfil its financial obligations the next day or its capital reserves fall below the required regulatory minimum.[6] The FDIC is required to resolve outstanding issues with problem banks in a manner that imposes the least cost on the deposit insurance fund (DIF) and ultimately on the taxpayer. Thus, early detection of insolvent institutions is of vital importance, especially if the failure of those institutions would pose a serious systemic risk on the financial system and the economy as a whole. The FDIC and state authorities utilize on-site and off-site examination methods in order to determine which institutions are insolvent and, thus, should be either closed or be provided financial assistance in order to rescue them. The off-site examinations are typically based on statistical and other mathematical methods and constitute complementary tools to the on-site visits made by supervisors to institutions considered at risk. There are three advantages to off-site versus on-site examinations: (i) the on-site examinations are more costly as they require the FDIC to bear the cost of visits and to retain extra staff during times when economic conditions are stable; (ii) the on-site examinations are usually time-consuming and cannot be performed with high frequency; and (iii) the off-site examinations can help allocate and coordinate the limited on-site examination resources in an efficient way with priority given to financial institutions facing the most severe challenges. The major drawback of statistically-based off-site tools is that they incorporate estimation errors which may affect the classification of banks as failure and nonfailures. An effective off-site examination tool must aim at identifying problem banks sufficiently prior to the time when a marked deterioration of their financial health would occur. Therefore, it is desirable to develop a model which would identify future failures with a high degree of accuracy in a timely manner and would rarely flag healthy banks as being at risk of closure.

This paper develops an early warning model of bank troubles and failures based on the Mixture Hazard Model (MHM) of Farewell (1977, 1982) with continuous and discrete time specifications.[7] MHM effectively combines the static model, which is used to identify troubled banks, and the duration model, which provides estimates of the probability of failure along with the timing of closure of such

[5]Community banks are banks with assets sizes of $1 billion or less. Their operation is oftentimes limited to rural communities and small cities. Community banks usually engage in traditional banking activities and provide more personal-based services.

[6]Under the current regulations issued by the Basel Committee on Banking Supervision (Basel II&III), a bank is considered as failed if its ratio of Tier 1 (core) capital to risk-weighted assets is 2% or lower. This ratio must exceed 4% to avoid supervisory intervention and prompt corrective action as underlined in Federal Deposit Insurance Corporation Improvement Act (FDICIA) of 1992. A bank with ratio of 6% or above is considered as a well-capitalized bank.

[7]Applications of the discrete-time version of the MHM can be found in Gonzalez-Hermosillo et al. (1997), Yildirim (2008) and Topaloglu and Yildirim (2009).

troubled banks. We view the financial crisis as a negative shock that affected banks in an unequal way. Well-capitalized, well-prepared, and prudently-managed institutions may have felt relatively little distress during the financial turmoil. On the other hand, poorly-managed banks that previously engaged in risky business practices faced an increased probability of their being on the FDIC watch list and, subsequently forced into closure or merger with a surviving bank by regulatory authorities. Unlike standard duration models, which assume that all banks are at the risk of failure, we implicitly assume that a proportion of banks will survive for a sufficiently long time after the end of a crisis and thus are incapable of entering an absorption state. In other words, we assume that the probability of failure for a bank that has never been on the watch list is arbitrarily close to zero. The MHM is appropriate for dealing with this issue as it is able to distinguish between healthy and at-risk of failure banks.

One of our (testable) assumptions concerns the fact that banks with low performance, as calculated by the radial measure of realized outcome to the maximum potential outcome, will increase their probability of failure. An inefficiently-managed bank could cumulatively save valuable funds by improving its performance. The saved funds often prove to be vital in servicing a bank's short-term obligations during financial crisis periods when interbank markets suffer from poor liquidity, and would therefore prevent the bank to need to draw on shareholders' equity. Shareholders' equity is the most expensive source of financing, the reduction of which would trigger on-site examination by regulators and possibly would place the bank on the watch list of problem banks. On-site examination subsequently would redirect the bank management's focus on clearing problem accounts rather than on improving its overall performance and thus could make it even less efficient. This process could continue in a spiral fashion, deteriorating the bank's financial assets and the capital. To account for this mutual effect, we employ a single step joint estimation procedure proposed by Tsionas and Papadogonas (2006), wherein a stochastic frontier model (SFM) is jointly estimated with a frailty model.

A challenge that we face in this paper is the incomplete information associated with the troubled banks on the watch list of the FDIC. Each quarter the FDIC releases the number of problem banks, but their identities are not publicly disclosed. To address this problem of missing information, we make an assumption that a bank that failed was on this list and based on available information we make a prediction of which banks potentially could be on this list through an expectation-maximization (EM) algorithm, which is designed to deal with this type of incomplete information. We also follow a forward step-wise procedure in model building and covariates selection, which is not only based on the conventional measures of the goodness-of-fit and statistical tests, but also on the contribution of these covariates to the predictive accuracy of the proposed models.

Finally, our model recognizes the fact that insolvency and failure are two different events. The realization of the first event is largely attributed to the actions undertaken by a bank itself, while the second usually occurs as a result of regulators' intervention following its insolvency. Supervisors typically tend not to seize

an insolvent bank unless it has no realistic probability of survival and its closure does not threaten the soundness and the stability of the financial system. Based on the above considerations, we are able to assess the type I and type II errors implicit in bank examiners' decision process when closing banks.[8] We find that the within sample and out-of-sample average of the two misclassification errors is less than 6 and 2%, respectively, for our preferred model. We also find that the predictive power of our model is quite robust when using estimates derived from different sub-periods of the financial crisis.

The remainder of the paper is organized as follows. In Sect. 14.2 we provide a brief review of banking crisis models. Section 14.3 describes the potential decision rule adopted by the regulatory authorities in determining and closing insolvent banks, which naturally will lead to the MHM. Two variants of the MHM are discussed, the continuous-time semiparametric proportional MHM and discrete-time MHM. In Sect. 14.4 we discuss the joint MHM-SFM. Section 14.5 deals with empirical specification issues and the data description. Estimation, testing, and predictive accuracy results are provided in Sect. 14.6, along with a comparison of various models and specifications. Section 14.7 contains our main conclusions.

14.2 Banking Crisis Studies

Accurate statistical models that serve as early warning tools and that potentially could be used as an alternative or complement to costly on-site visits made by supervisors have been well documented in the banking literature. These models have been applied successfully to study banking and other financial institutions' failures in the U.S. and in other countries. As the literature that deals with bankruptcy prediction of financial institutions is vast and there are a myriad of papers that specifically refer to the banking industry failures, we will discuss only few the studies that are closely related to our work and are viewed as early warning models.

The more widely-used statistical models for bankruptcy prediction are the single-period static probit/logit models and the methods of discriminant analysis. These method usually estimate the probability that an entity with specific characteristics will fail or survive within a certain time interval. The timing of the failure is not provided by such models. Applications of the probit/logit models and discriminant analysis can be found in Altman (1968), Meyer and Pifer (1970), Deakin (1972), Martin (1977), Lane et al. (1986), Cole and Gunther (1995, 1998), Cole and Wu (2011), among others.

Others in this literature have employed the Cox proportional hazard model (PHM) and the discrete time hazard model (DTHM) to explain banking failures and

[8]Typically, a type I error is defined as the error due to classifying a failed bank as a non-failed bank, while a type II error arises from classifying a non-failed bank as a failed bank.

develop early warning models.[9] In the hazard model the dependent variable is time to the occurrence of some specific event, which can be equivalently expressed either through the probability distribution function or the hazard function, which provides the instantaneous risk of failure at some specific time conditional on the survival up to this time. The PHM has three advantages over the static probit/logit models: (i) it provides not only the measure of probability of failure (survival), but also the probable timing of failure; (ii) it accommodates censored observations, those observations that survive through the end of the sample period; and (iii) it does not make strong assumptions about the distribution of duration times. The disadvantage of the PHM model is that it requires the hazard rate to be proportional to the baseline hazard between any two cross-sectional observations. Moreover, inclusion of time-varying covariates is problematic. The DTHM, on the other hand, easily allows for time-varying covariates and has the potential to provide more efficient estimates and improved predictions. The application of PHM to the study the U.S. commercial banking failures was undertaken by Lane et al. (1986), Whalen (1991), as well as in Wheelock and Wilson (1995, 2000).[10]

Barr and Siems (1994) and Wheelock and Wilson (1995, 2000) were the first to consider inefficiency as a potential influential factor explaining U.S. commercial banking failures during the earlier crisis. They estimated the efficiency scores with Data Envelopment Analysis (DEA) techniques, which were then used in a static model to predict banking failures.[11] Wheelock and Wilson (1995, 2000), on the hand, included inefficiency scores among their regressors to allow these to affect the probability of failure and acquisitions by other banks in the PHM framework. They employed three measures of radial technical inefficiency, namely the parametric cost inefficiency measure, the nonparametric input distance function measure, and the inverse of the nonparametric output distance function measure. According to the authors, the first two had a statistically significant (positive) effect on the probability of failure, while only the first measure significantly decreased the acquisition probability. The estimation of the models was conducted in two stages. The first stage involved the parametric or nonparametric estimation of inefficiency scores. In the second stage these scores were used among the explanatory variables to investigate their effect on the probabilities of failure and acquisition. Tsionas and Papadogonas (2006) criticize the two-step approach, arguing that this may entail an error-in-variables bias as well as introduce an endogenous auxiliary regressor.

[9]A thorough discussion of hazard models can be found in Cox (1972), Lancaster (1990), Kalbfleisch and Prentice (2002), and Klein and Moeschberger (2003).

[10]Shumway (2001), Halling and Hayden (2006), Cole and Wu (2009), and Torna (2010) provide non-banking applications, along with arguments for using the DTHM over the static models and PHM.

[11]DEA, which was proposed by Charnes et al. (1978), is a nonparametric approach that estimates a relative efficiency score for a bank based on linear programming techniques.

14.3 Mixture Hazard Model

Our banking failure modelling approach is based on the rules and policies that regulatory authorities implement in order to identify problem banks that subsequently fail or survive.[12] We first let H_{it} define the financial health indicator of bank i at time t and assume that there is a threshold level of it, H_{it}^*, such that if the financial health falls bellow this level then the bank is considered at risk of closure by regulatory authorities. Second, we let the difference between H_{it}^* and H_{it}, denoted by h_{it}^*, be dependent on bank-specific financial metrics and market variables as follows:

$$h_{it}^* = H_{it}^* - H_{it} = x_{it}'\boldsymbol{\beta} + e_{it} \tag{14.1}$$

where e_{it} represents the error term, which is assumed to be identically and independently distributed (*iid*) across observations and over time.[13]

The financial health threshold of a particular bank is a composite and oftentimes subjective index and its lower bound is not observable to the econometrician; therefore, h_{it}^* is not observable as well. Instead a binary variable h_{it} can be defined such that

$$h_{it} = \begin{cases} 1 & \text{if } h_{it}^* > 0 \\ 0 & \text{if } h_{it}^* \leq 0 \end{cases}$$

Based on the above, the probability that a bank will become a *problem* bank is given by

$$P(h_{it} = 1) = P(h_{it}^* > 0) = P(e_{it} > -x_{it}'\boldsymbol{\beta}) = F_e(x_{it}'\boldsymbol{\beta})$$

where F_e is the cumulative distribution function of the random error e, which can be assumed to be either normally distributed (probit model) or logistically distributed (logit model).

Specification of the likelihood function then follows that of the standard hazard model, wherein a nonnegative random variable T represents the time to failure of a bank within a given period of time. This is characterized by the conditional probability density function f_T and the cumulative distribution function F_T. A binary variable d_i is also specified and takes on a value of 1 for observations that fail at time t and 0 for observations that are right censored (i.e., when a bank does not fail by the end of the sample period or disappears during the period for reasons

[12] See Kasa and Spiegel (2008) on various regulatory closure rules.

[13] The *iid* assumption of the error term can be relaxed in the panel data context by assuming $e_{it} = \mu_i + \xi_{it}$ with $\mu_i \sim N(0, \sigma_\mu^2)$ and $\xi_{it} \sim N(0, \sigma_\xi^2)$ independent of each other. This adds an additional complication to the model and it is not pursued in this paper.

other than failure).[14] Assuming that the rate at which regulatory authorities tend to seize healthy banks is arbitrarily close to zero, the likelihood function for a bank i is given by

$$L_i(\boldsymbol{\theta}; x, w) = \left[F_e(x_i'\boldsymbol{\beta})\lambda_i^p(t; w_i)S^p(t; w_i)\right]^{d_i}\left(F_e(x_i'\boldsymbol{\beta})S^p(t; w_i) + \left[1 - F_e(x_i'\boldsymbol{\beta})\right]\right)^{1-d_i} \tag{14.2}$$

where S^p is a survivor function, which represents the probability that a problem bank will survive for a period longer than t and λ^p represents the hazard rate or probability that such bank will fail during the next instant, given that it was in operation up until this time. The $\boldsymbol{\theta}$ represents the parameter vector, while x and w are covariates associated with the probability of being problem and failed, respectively. A detailed derivation of the likelihood function is provided in Appendix A of this paper. After rearranging the expression in (14.2) and dropping the superscript from measures pertaining to problem banks to reduce notational clutter, the sample likelihood for all banks can be written as:

$$\begin{aligned} L(\boldsymbol{\theta}; x, w, d) &= \prod_{i=1}^{n} L_i(\boldsymbol{\theta}; x, w, d) \\ &= \prod_{i=1}^{n} F_e(x_i\boldsymbol{\beta})^{h_{it}}(1 - F_e(x_i\boldsymbol{\beta}))^{1-h_{it}}\left\{\lambda_i(t; w_i)\right\}^{d_i h_{it}}\left\{S_i(t; w_i)\right\}^{h_{it}} \end{aligned} \tag{14.3}$$

If T is assumed to be a time-varying variable, then the model can be estimated based on the proportional hazards assumption (Cox 1972), which unfortunately does not allow for time-varying covariates. Following Kuk and Chen (1992) and Sy and Taylor (2000), the survivor and hazard functions in this case are given by the expressions bellow:

$$\lambda_i(t; w_i) = \lambda_0(t)\exp\left(w_i'\boldsymbol{\alpha}\right) \text{ and } S_i(t; w_i) = S_0(t)^{\exp(w_i'\boldsymbol{\alpha})} \tag{14.4}$$

where, $\lambda_0(t)$ and $S_0(t)$ define the conditional baseline hazard function and baseline survivor function, respectively. These are nonnegative functions of time only and are assumed to be common to all banks at risk.

The discrete-time version of the model on the other hand is more flexible and adds more dynamics to the model by allowing for inclusion of time-varying covariates. This specification, however, requires that the time-varying regressors

[14]In this paper, failed banks are only considered as the banks that appear on the FDIC's failed bank list. Banks that ceased their operation due to reasons other than failure (e.g., merger or voluntary liquidation) or remained inactive or are no longer regulated by the Federal Reserve, have censored duration times.

remain unchanged in the time interval $[t, t+1]$. The survivor and hazard functions in the discrete-time MHM can be shown to be derived as:[15]

$$S_{ij}(t; w, u) = \left[\prod_{j=1}^{t_i} \frac{1}{1+\exp(w'_{ij}\boldsymbol{\alpha})}\right] \text{and } \lambda_{ij}(t; w) = 1 - \frac{S(t_{ij})}{S(t_{i,j-1})} for\, j = 1, 2, \ldots, t_i.$$

In what follows, we refer to the continuous-time MHM as Model I and the discrete-time MHM as Model II. Following the standard nomenclature in the medical and biological sciences, we refer to the portion of the model that assesses the financial health of a bank as the **incidence** component and the portion of the model that assesses survival times as the **latency** component.

If h_{it} is observed by the econometrician for each individual bank as it is by regulators then the estimation process reduces to that of the standard MHM. However, as discussed above h_{it} is only partially observed by the econometrician. We address this problem of incomplete information by utilizing the EM algorithm to deal with the missing data. The EM algorithm consists of two iterative steps: the expectation (E) step and the maximization (M) step. The expectation step involves the projection of an appropriate functional (likelihood or log-likelihood function) containing the augmented data on the space of the original (incomplete) data. Thus, the missing data are first estimated, given the observed data and the initial estimates of the model parameters, in the E step. In the M step the function is maximized while treating the incomplete data as known. Iterating between these two steps yields estimates that under suitable regulatory conditions converge to the maximum likelihood estimates (MLE).[16]

To implement the EM algorithm first consider the expectation of the full log-likelihood function with the respect to the h_{it} and the data, which completes the E step of the algorithm. Linearity of log-likelihood function with respect to the h_{it} considerably facilitates the calculations and the analysis.

The log-likelihood for the i^{th} observation in the M step is given by:

$$\begin{aligned} E^{(M)}_{h|X,W,\theta,\lambda_0}[L_i(\boldsymbol{\theta}; x, w, d)] = {} & \tilde{h}^{(M)}_{it} \log[F_e(x_i\boldsymbol{\beta})] + (1 - \tilde{h}^{(M)}_{it}) \log[1 - F_e(x_i\boldsymbol{\beta})] \\ & + \tilde{h}^{(M)}_{it} d_i \log[\lambda_i(t; w_i)] + \tilde{h}^{(M)}_{it} \log[S_i(t; w_i)] \end{aligned} \tag{14.5}$$

where $\tilde{h}_{it}$ is the probability that the ith bank will eventually belong to the group of problem banks at time t, conditioned on the observed data and the model parameters. It represents the fractional allocation to the problem banks and is given by:

[15]See Cox and Oaks (1984), Kalbfleisch and Prentice (2002), and Bover et al. (2002) for discussion on discrete-time proportional hazard models.

[16]For more discussion on the EM algorithm and its convergence properties and limitations see Dempster et al. (1977) as well as McLachlan and Krishnan (1996).

$$\tilde{h}_{it}^{(M)} = E\left[h_{it}|\boldsymbol{\theta}^{(M)}, Data\right] = \Pr(h_{it}^{(M)} = 1|t_i > T_i)$$
$$= \begin{cases} \frac{F_e(x_i'\boldsymbol{\beta}^{(M)})S_i(t;w_i)}{F_e(x_i'\boldsymbol{\beta}^{(M)})S_i(t;w_i) + (1-F_e(x_i'\boldsymbol{\beta}^{(M)}))} & \text{if } d_i = 0 \\ 1 \quad \text{otherwise} \end{cases} \tag{14.6}$$

In Model I, the nuisance baseline hazard function λ_0 is estimated nonparametrically from the profile likelihood function as:

$$\hat{\lambda}_0(t) = \frac{N(t_i)}{\sum_{j\in R(t_i)} \tilde{h}_{jt} \exp(w_j'\boldsymbol{\alpha})} \tag{14.7}$$

where $N(t_i)$ is the number of failures and $R(t_i)$ is the set of all individuals at risk at time t_i, respectively. Substituting (14.7) into (14.5) leads to the M step log-likelihood for Model I:

$$\begin{aligned} \tilde{L}(\boldsymbol{\theta}; x, w, \tilde{h}) &= \sum_{i=1}^{n} \left\{\tilde{h}_{it} \log F_e(x_i'\boldsymbol{\beta}) + (1-\tilde{h}_{it}) \log(1 - F_e(x_i'\boldsymbol{\beta}))\right\} \\ &+ \sum_{i=1}^{N} \left\{w_i'\boldsymbol{\alpha} - N(t_i) \log\left(\sum\nolimits_{j\in R(t_i)} \tilde{h}_{jt} \exp(w_j'\boldsymbol{\alpha})\right)\right\} \\ &= L_1(\boldsymbol{\beta}; x, \tilde{h}) + \tilde{L}_2(\boldsymbol{\alpha}; w, \tilde{h}) \end{aligned} \tag{14.8}$$

The full implementation of the EM algorithm involves the following four steps:

- Step 1: Provide an initial estimate for the parameter $\boldsymbol{\beta}$ and estimate the ordinary MHM in order to obtain the starting values for λ_0;
- Step 2 (E step): Compute $\tilde{h}_{it}$ from (14.6) based on the current estimates and the observed data;
- Step 3 (M step): Update the estimate of parameter $\boldsymbol{\beta}$ using (14.5); and
- Step 4: Iterate between steps 2 and 3 until convergence is reached.[17]

Alternatives to the EM method can also be utilized. For example, in his study of the recent U.S. commercial banking failures Torna (2010) attempted to identify troubled banks on the FDIC's watch list based on their tier 1 capital ranking. Banks were ranked according to their tier 1 capital ratio and the number of banks with the lowest value were selected to match the number provided by the FDIC in each quarter. Other ratios, such as Texas ratio, also can be utilized to deduce the problem banks. The Texas ratio was developed by Gerard Cassidy to predict banking

[17]Convergence to a stationary point in the EM algorithm is guaranteed since the algorithm aims at increasing the log-likelihood function at each iteration stage. The stationary point need not, however, be a local maximum. It is possible for the algorithm to converge to local maxima or saddle points. We check for these possibilities by selecting different starting values and checking for the proper signs of the Hessian.

failures in Texas and New England during recessionary periods of the 1980s and 1990s. It is defined as the ratio of nonperforming assets to total equity and loan-loss reserves and banks with ratios close to one are identified as high risk. There are at least two limitations to these approaches besides their crude approximation. First, they ignore other variables that play a pivotal role in leading banks to a distressed state. For example, ratios based on nonperforming loans are major indicators of difficulties that bank will face in near future, even if their current capital ratios are at normal levels. Second, financial ratios that are used to classify banks as healthy or troubled cannot be subsequently employed as determinants due to a possible endogeneity problem.

14.3.1 Combined SFM and MHM Model

In this section we consider the efficiency performance of a bank as a determinant of the probability of being both a problem bank and one that subsequently fails. The efficiency performance of a firm relative to the best practice (frontier) technology firm was formally considered by Debreu (1951) and Farrell (1957). Aigner et al. (1977), Meeusen and van den Broeck (1977), and Battese and Cora (1977) introduced the parametric stochastic frontier model (SFM). In the SFM the error term is assumed to be multiplicative in a levels specification of the production or of one of its dual presentations, such as the cost function we use in our analysis, and is composed of two parts: (i) a one-sided error term that captures the effects of inefficiencies relative to the stochastic frontier; and (ii) a two-sided error term that captures random shocks, measurement errors and other statistical noise.[18]

The general SFM is represented by the following functional relationship:

$$y_{it} = g(z_{it}; \boldsymbol{\eta}) \exp(\varepsilon_{it}) \tag{14.9}$$

where the dependent variable y_{it} could represent cost, output, profit, revenue and so forth, z_{it} is a vector of independent regressors, and $g(\cdot)$ is the frontier function, which can be either linear or non-linear in coefficients and covariates. Depending on the particular dual representation of technology specified, $\varepsilon = \upsilon \pm u\,[= \log y_{it} - \log\, g(z_{it}; \boldsymbol{\eta})]$ represents the composed error term, with υ_{it} representing the noise and u_i the inefficiency process. The noise term is assumed to be *iid* normally distributed with zero mean and constant variance. Inefficiencies are also assumed to be *iid* random variables with distribution function defined on the domain of positive numbers $(u \in R_+)$. Both υ and u are assumed to be independent

[18]Excellent surveys of frontier models and their applications are found in Kumbhakar and Lovell (2000) and Greene (2008).

of each other and independent of the regressors.[19] In this paper, we follow Pitt and Lee (1981) and assume that the inefficiency process is a time-invariant random effect, which follows the half-normal distribution (i.e., $u_i \sim N^+(0, \sigma_u^2)$).

Under the above assumptions the marginal distribution of the composed error term, which for the production or cost frontier model is derived as:

$$f_\varepsilon(\varepsilon_{it}) = \frac{2}{(2\pi)^{T_i/2}\sigma_\upsilon^{T_i-1}\sigma}\exp\left[-\frac{\varepsilon_{it}'\varepsilon_{it}}{2\sigma_\upsilon^2} + \frac{\bar{\varepsilon}_i^2\lambda^2}{2\sigma^2}\right]\left[1-\Phi\left(\frac{T_i\bar{\varepsilon}_i\lambda}{\sigma}\right)\right] \tag{14.10}$$

where $\sigma = \sqrt{\sigma_\upsilon^2 + T_i\sigma_u^2}$, $\lambda = \sigma_u/\sigma_\upsilon$, and $\bar{\varepsilon}_i = (1/T_i)\sum_{t=1}^{T_i}\varepsilon_{it}$.[20] The parameter λ is the signal-to-noise ratio and measures the relative allocation of total variation to the inefficiency term. In practice we can use an alternative parametrization, called the γ—parameterization, which specifies $\gamma = \sigma_u^2/\sigma^2$.[21]

It can be also shown (see Jondrow et al. 1982) that the conditional distribution of the inefficiency term is given by

$$f_{u|\varepsilon}(u_i|\varepsilon_{it}) = \frac{f_{\varepsilon,u}(\varepsilon_i, u_i)}{f_\varepsilon(\varepsilon_i)} = \frac{\frac{1}{\sigma}\phi\left(\frac{u_i-\mu_i^*}{\sigma_*}\right)}{\left[1-\Phi\left(-\frac{\mu_i^*}{\sigma_*}\right)\right]} \tag{14.11}$$

where $f_{u|\varepsilon}(\cdot)$ represents the normal distribution truncated at 0 with mean $\mu_i^* = -T_i\bar{\varepsilon}_i\sigma_u^2/\sigma^2 = -T_i\bar{\varepsilon}_i\gamma$ and variance $\sigma_*^2 = \sigma_u^2\sigma_\upsilon^2/\sigma^2 = \gamma\sigma^2(1-\gamma T_i)$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the pdf and cdf functions of the standard normal distribution. The mean or the mode of this conditional distribution function provides an estimate of the technical inefficiency.

In the absence of any effect of the inefficiencies on the probability of being troubled and failed, (14.10) and (14.11) can be employed to obtain the maximum likelihood estimates of model parameters and efficiency scores. However, consistent and efficient parameter estimates cannot be based solely on the frontier model when there is feedback between this measure of economic frailty and the likelihood of failure and the ensuing tightening of regulatory supervision. There is a clear need for joint estimation of the system when the decision of a firm is affected by these factors.

---

[19]The assumption of independence of the inefficiency term and the regressors is restrictive, but is necessary for our current analysis. Its validity can be tested using the Hausman-Wu test. In the panel data context, this assumption can be relaxed by assuming that inefficiencies are fixed effects or random effects correlated with all or some of the regressors (Hausman and Taylor 1981; Cornwell et al. 1990).

[20]The cost frontier is obtained by reversing the sign of the composed error.

[21]This reparametrization is desirable as the γ parameter has compact support, which facilitates the numerical procedure of maximum likelihood estimation, hypothesis testing, and establishing the asymptotic normality of this parameter.

In deriving the likelihood function for this model, we maintain the assumption that censoring is non-informative and statistically independent of h_i. Following Tsionas and Papadogonas (2006) we also assume that conditional on inefficiency and the data the censoring mechanism and h_i are independent of the composed error term. To simplify notations, let $\boldsymbol{\Omega}_i = \{x_i, w_i, z_i\}$ denote the set of covariates and $\Theta = \left\{\boldsymbol{\beta}, \boldsymbol{\alpha}, \delta_1, \delta_2, \boldsymbol{\eta}, \sigma_v^2, \sigma_u^2\right\}$ be the vector of the structural and distributional parameters. The observed joint density of the structural model for bank i, given h_i and after integrating out the unobserved inefficiency term, can be written as:

$$\begin{aligned} L_i(y_i, h_i, d_i|\boldsymbol{\Omega}_i, \Theta') &= \int_0^{\infty} F_e(x_i'\boldsymbol{\beta} + \delta_1 u_i)^{h_i}(1 - F_e(x_i'\boldsymbol{\beta} + \delta_1 u_i))^{1-h_i} \\ &\quad \times \{\lambda_i(t; w_i, u_i)\}^{d_i h_i}\{S_i(t; w_i, u_i)\}^{h_i} \underbrace{f_v(\varepsilon_{it} \pm u_i) f(u_i)}_{f_\varepsilon(\varepsilon) f_{u|\varepsilon}(u|\varepsilon)} du_i \\ &= f_\varepsilon(\varepsilon_{it}) \int_0^{\infty} F_e(x_i'\boldsymbol{\beta} + \delta_1 u_i)^{h_i}(1 - F_e(x_i'\boldsymbol{\beta} + \delta_1 u_i))^{1-h_i} \\ &\quad \times \{\lambda_i(t; w_i, u_i)\}^{d_i h_i}\{S_i(t; w_i, u_i)\}^{h_i} f_{u|\varepsilon}(u|\varepsilon) du_i. \end{aligned} \tag{14.12}$$

The hazard rate and survival function for the continuous-time counterpart of the model are now given by:

$$\lambda_i(t; w_i, u_i) = \lambda_0(t)\ \exp\left(w_i'\boldsymbol{\alpha} + \delta_2 u_i\right) \text{ and } S(t; w_i) = S_0(t)^{\exp\left(w_i'\boldsymbol{\alpha} + \delta_2 u_i\right)}$$

It should be noted that the above model is rather a general one and consists of three individual parts: (1) the SFM; (2) the probit/logit model for the incidence part; and (3) the standard hazard model for the latency part. Each of these three models are nested within the general model. For example, if there is no association between inefficiency and the probability of being troubled or failed ($\delta_1 = 0$ and $\delta_2 = 0$), then (14.12) consists of two distinct parts, the SFM and the MHM. Both can be estimated separately using the methods outlined in the previous sections.

The integral in the joint likelihood (14.12) has no closed form solution and thus the maximization of this function requires numerical techniques, such as simulated maximum likelihood (SML) or Gaussian quadrature.[22] In SML the sample of draws from $f_{u|\varepsilon}(\cdot)$ are required to approximate the integral by its numerical average (expectation). As such, the simulated log-likelihood function for the ith observation becomes:

[22]Tsionas and Papadogonas (2006) employed the Gaussian quadrature in estimation of the model where the technical inefficiency has a potential effect on firm exit. Sickles and Taubman (1986) used similar methods in specifying structural models of latent health and retirement status, while controlling for multivariate unobserved individual heterogeneity in the retirement decision and in morbidity.

$$L_i = \log L_i(y_i, h_i, d_i|\boldsymbol{\Omega}_i, \boldsymbol{\Theta}') = Constant - \frac{(T_i - 1)}{2}\log\sigma^2(1-\gamma T_i)$$
$$-\frac{1}{2}\log\sigma^2 + \log\left[1 - \Phi\left(\frac{T_i\bar{\varepsilon}_i\sqrt{\gamma/(1-\gamma)}}{\sigma}\right)\right] - \frac{\varepsilon_{it}'\varepsilon_{it}}{2\sigma^2(1-\gamma T_i)} + \frac{\bar{\varepsilon}_i^2\gamma}{2\sigma^2(1-\gamma)}$$
$$+ \log\frac{1}{J}\sum_{j=1}^{J}\left\{F_e(x_i\boldsymbol{\beta} + \delta_1 u_{ij})^{h_{it}}\left[1 - F_e(x_i\boldsymbol{\beta} + \delta_1 u_{ij})\right]^{1-h_{it}}\left[\lambda_i(t; w_i, u_{ij})\right]^{d_i h_{it}}\left[S_i(t; w_i, u_{ij})\right]^{h_{it}}\right\} \tag{14.13}$$

where u_{is} is a random draw from the truncated normal distribution $f_{u|\varepsilon}(\cdot)$ and J is the number of draws. We utilize the inverse cdf method to efficiently obtain draws from this distribution as:

$$u_{ij} = \mu_i^* + \sigma_*\Phi^{-1}\left[U_{is} + (1 - U_{is})\Phi\left(-\frac{\mu_i^*}{\sigma_*}\right)\right] \tag{14.14}$$

where U is a random draw from uniform $U[0, 1]$ distribution or a Halton draw. By substituting (14.14) into (14.13) and treating the $h_{it}s$ as known we can maximize the log-likelihood function $L = \sum_i L_i$ by employing standard optimization techniques and obtain the model parameters.

Finally, after estimating the model parameters, the efficiency scores are obtained as the expected values of the conditional distribution, in the spirit of Jondrow et al. (1982):

$$\hat{u}_i = E\left[u_i|\hat{\varepsilon}_i, \tilde{h}_i, d_i, \boldsymbol{\Omega}_i, \boldsymbol{\Theta}'\right] = \frac{\int_0^\infty u_i G(u_i; \boldsymbol{\Theta}) f_{u|\varepsilon}(u|\varepsilon)du_i}{\int_0^\infty G(u_i; \boldsymbol{\Theta}) f_{u|\varepsilon}(u|\varepsilon)du_i} \tag{14.15}$$

where $G(u_i; \Theta) = \tilde{F}(x_i'\boldsymbol{\beta} + \delta_1 u_i)^{\tilde{h}_{it}}\left[1 - \tilde{F}(x_i'\boldsymbol{\beta} + \delta_1 u_i)\right]^{1-\tilde{h}_{it}}$ $\left[\lambda_i(t; w_i, u_i)\right]^{d_i\tilde{h}_{it}}$ $\left[S(t; w_i, u_i)\right]^{\tilde{h}_{it}}$. The integrals in the numerator and denominator are calculated numerically by the SML method and the efficiency score of the ith firm is estimated as $TE_i = \exp(-\hat{u}_i)$. It is straightforward to check that if δ is zero then (14.15) collapses to the formula derived by Jondrow et al. for production frontiers (i.e., $\hat{u}_i = E[u_i|\hat{\varepsilon}_i] = \mu_* + \sigma\phi\left(\frac{\mu_*}{\sigma_*}\right)/\Phi\left(\frac{\mu_*}{\sigma_*}\right)$).

The EM algorithm for the stochastic frontier MHM involves the following steps:

- Step 1: Provide initial estimates of the parameter vector $\boldsymbol{\Theta}$. Set the initial value of parameters δ_1 and δ_2 equal to zero and obtain the initial value of the baseline hazard function from (14.7). Consistent starting values of the variances of the noise and inefficiency terms are based on method of moments estimates

$$\begin{aligned}\hat{\sigma}_u^2 &= \left[\sqrt{2/\pi}\left(\frac{\pi}{\pi-4}\right)\hat{m}_3\right]^{2/3} \\ \hat{\sigma}_v^2 &= \left[\hat{m}_2 - \left(\frac{\pi-2}{\pi}\right)\hat{\sigma}_u^2\right]\end{aligned} \tag{14.16}$$

where $\hat{m}_2$ and $\hat{m}_3$ are the estimated second and third sample moments of the OLS residuals, respectively. Estimates of σ and γ parameters are obtained through the relevant expressions provided above.

- Step 2 (E step): Compute $\tilde{h}_{it}$ based on the current estimates and the observed data from

$$\tilde{h}_{it}^{(M)} = E\left[h_{it}|\boldsymbol{\Theta}^{(M)}, \boldsymbol{\Omega}_i\right] = \Pr(h_{it}^{(M)} = 1|t_i > T_i)$$
$$\begin{cases} \frac{F_e(x_i'\boldsymbol{\beta}^{(M)} + \delta_1^{(M)}u_i)S_i(t;w_i,u_i)}{F_e(x_i'\boldsymbol{\beta}^{(M)} + \delta_1^{(M)}u_i)S_i(t;w_i,u_i) + (1-F_e(x_i'\boldsymbol{\beta}^{(M)} + \delta_1^{(M)}u_i))} & \text{if } d_i = 0 \\ 1 & \text{otherwise} \end{cases} \tag{14.17}$$

- Step 3 (M step): Update the estimate of parameters by maximizing L via simulated maximum likelihood technique.
- Step 4: Iterate between steps 2 and 3 until convergence.
  Continuous-time and discreet-time versions of this combined model are referred as Model III and Model IV, respectively, throughout this paper.

14.4 Empirical Model and Data

In this section we outline the empirical specification used in estimating the four models described above (Models I–IV). We describe the data on which our estimates are based and the step-wise forward selection procedure we employ in model building and variable selection.

14.4.1 Empirical Specification

Following Whalen (1991) we employ a model with a two-year timeline to estimate the probability of distress and failure and the *timing* of bank failure. In the Model I and Model III, the time to failure is measured in months[23] (1–24 months) starting from the end-year of 2007, while in the Model II and Model IV the duration times are measured in quarters as banks report their data on a quarterly basis. The

[23]Duration times measured in weeks were also considered, but not reported in this paper.

covariates used in the estimation process of Model I and Model III are based on information from the fourth quarter of the 2007 Consolidated Reports of Condition and Income (Call Reports). State-specific macroeconomic variables are also derived from the Federal Reserve databases to control for state-specific effects.

We employ the cost frontier in the stochastic frontier model specification, which describes the minimum level of cost given output and input prices. The gap between the actual cost and the minimum cost is the radial measure of total (cost) inefficiency and is composed of two parts: (i) the technical inefficiency arising from excess usage of inputs and (ii) the allocative inefficiency that results from a non-optimal mix of inputs. We do not make this decomposition but rather estimate overall cost inefficiency. We adopt the intermediation approach of Sealey and Lindley (1977), according to which banks are viewed as financial intermediaries that collect deposits and other funds and transform them into loanable funds by using capital and labor. Deposits are viewed as inputs as opposed to outputs, which is assumed in the production approach.[24]

As in Kaparakis et al. (1994) and Wheelock and Wilson (1995), we specify a multiple output-input short-run stochastic cost frontier with a quasi-fixed input. Following the standard banking literature we specify a translog functional form to describe the cost function:[25]

$$\begin{aligned}\log C_{it} = \alpha_0 + & \sum_{m=1}^{5} \alpha_m \log y_{mit} + \sum_{k=1}^{4} \beta_k \log w_{kit} \\ & + \frac{1}{2}\sum_{m=1}^{5}\sum_{j=1}^{5} \alpha_{mj} \log y_{mit} \log y_{jit} + \theta_1 t + \frac{1}{2}\theta_2 t^2 \\ & + \frac{1}{2}\sum_{k=1}^{4}\sum_{n=1}^{4} \beta_{kn} \log w_{kit} \log w_{nit} + \eta_1 \log X_{it} + \frac{1}{2}\eta_2 (\log X_{it})^2 \\ & + \sum_{m=1}^{5}\sum_{k=1}^{4} \delta_{mk} \log y_{mit} \log w_{kit} + \sum_{m=1}^{5} \lambda_{1x} \log y_{mit} \log X_{it} \\ & \sum_{k=1}^{4} \lambda_{2x} \log w_{kit} \log X_{it} + \sum_{m=1}^{5} \lambda_{mt} \log y_{mit} t + \sum_{k=1}^{4} \phi_{kt} \log w_{kit} t + \upsilon_{it} + u_i\end{aligned}$$

where C is the observed short-run variable cost of an individual bank at each time period, y_m is the value of the mth output, $m = 1, \ldots, 5$. Outputs are real estate loans $(yreln)$, commercial and industrial loans $(yciln)$, installment loans $(yinln)$, securities $(ysec)$, and off-balance sheet items $(yobs)$. The $w's$ represent input prices of the total interest-bearing deposits (dep), labor (lab), purchased funds $(purf)$, and

[24]See Baltensperger (1980) for example.

[25]The translog function provides a second-order differential approximation to an arbitrary function at a single point. It does not restrict the share of a particular input to be constant over time and across individual firms.

capital (cap). The quasi-fixed input (X) consists of non-interest-bearing deposits. Kaparakis et al. assume that a bank takes the level of non-interest-bearing deposits as exogenously given and since there is no market price associated with this input, the quantity of it should be included in the cost function instead of its price. We also include the time and its interaction with outputs and input prices to account for non-neutral technological change. Symmetry ($\alpha_{mj} = \alpha_{jm}$ and $\beta_{kn} = \beta_{nk}$) and linear homogeneity in input price ($\sum_{k=1}^{4} \beta_k = 1, \sum_{k=1}^{4} \beta_{kn} = \sum_{k=1}^{4} \delta_{mk} = \sum_{k=1}^{4} \lambda_{2x} = \sum_{k=1}^{4} \phi_{kt} = 0$) restrictions are imposed by considering capital as the numeraire and dividing the total cost and other input prices by its price.

14.4.2 Data

The data are from three main sources: (i) the public-use quarterly Call Reports for all U.S. commercial banks, which is collected and administrated by the Federal Reserve Bank of Chicago and the FDIC; (ii) the FDIC website, which provides information regarding failed banks and industry-level indicators; and (iii) the website of the Federal Reserve Bank of St. Louis, which provides information on regional-specific macroeconomic variables.

We drop bank observations with zero costs, zero output and input levels, as well as those with obvious measurement errors and other data inconsistencies. In addition, we exclude banks that voluntarily liquidated during the sample period and those that were chartered and started to report their data after the first quarter of 2007,[26] which require a special treatment. The estimation sample consists of 125 banks that failed during 2008 and 2009 and 5843 surviving banks.

More than forty bank-specific financial metrics, state-specific macroeconomic, geographical, and market structure variables are constructed from variables obtained form the above sources as potential determinants of banking distress and failure. We apply the stepwise forward selection procedure (Klein and Moeschberger 2003) to chose the most relevant explanatory variables based on conventional statistical tests and the Akaike Information Criterion (AIC). In addition to these tests, we base our variable selection on their contribution to the overall prediction accuracy of a particular model we employ. The final set of variables pertaining to the incidence and the latency part includes proxies for the capital adequacy, asset quality, management, earnings, liquidity, and sensitivity (the so-called "CAMELS"),[27] six market structure and geographical variables, and four state-specific variables. We use the same set of explanatory variables in both the incidence and latency parts of our models in order to capture the different effects

[26]These are typically referred to as the "De Novo" banks (DeYoung 1999, 2003).

[27]The "CAMELS" variables construction closely follows that of Lane et al. (1986) and Whalen (1991).

Table 14.1 CAMELS proxy financial ratios

Capital adequacy (C)	
tier1	Tier 1 (core) capital/risk-weighted assets
Asset quality (A)	
rnpl	Nonperforming loans/total loans
alll	Allowance for loan and lease loss/average loans and leases
reln	Commercial real estate loans/total loans
coffs	Charge-offs on loans and leases/average loans and leases
lrec	Recoveries on allowance for loan and lease losses/average loans and leases
llp	Provision for loan and lease losses/average loans and leases
Managerial quality (M)	
fte	(Number of fulltime equivalent employees/average assets) * 1000
imr	Total loans/total deposits
u	Random effects inefficiency score
Earnings (E)	
oi	Total operating income/average assets
roa	Net income (loss)/average assets
roe	Net income (loss)/total equity
Liquidity (L)	
cash	Noninterest-bearing balances, currency, and coin/average assets
cd	Total time deposits of US$100,000 or more/total assets
coredep	Core deposits/total assets
Sensitivity (S)	
sens	Difference in interest rate sensitive assets and liabilities repricing within one year/total assets

that these have on the probability that a particular bank is troubled, as well as the probability and timing of the resolution of the bank's troubles by the FDIC. Tables 14.1 and 14.2 provide our mnemonics for the variable names, as well as their formal definitions.

The first variable in Table 14.1 is the tier 1 risk-based capital ratio. Banks with a high level of this ratio are considered having sufficient capital to absorb the unexpected losses occurring during the crisis and hence, have a higher chance of survival. We expect a negative sign for this variable in both the incidence and latency parts. The next variable is the ratio of nonperforming loans[28] to total loans, which is the primary indicator of the quality of loans made by banks and historically has been an influential factor in explaining their distress and failure. The higher this ratio, the higher the probability that the bank will enter the watch list and subsequently fail. The next five ratios also reflect the banks' asset quality. We expect the ratio of allowance for loan and lease loss to average total loans to have a

[28]Nonperforming loans consist of total loans and lease financing receivables that are nonacrual, past due 30–89 days and still accruing, and past due 90 days or more and still accruing.

Table 14.2 Geographical, market structure, and state-specific macroeconomic variables

Geographical and market structure variables	
chtype	Charter type (1 if state chartered, 0 otherwise)
frsmb	FRS membership indicator (1 if federal reserve member, 0 otherwise)
ibf	International banking facility (1 if bank operates an international based facility, 0 otherwise)
frsdistrcode	FRS district code: [Boston(1), New York (2), Philadelphia (3), Cleveland (4), Richmond (5), Atlanta (6), Chicago (7), St. Louis (8), Minneapolis (9), Kansas City (10), Dallas (11), San Francisco (12), Washington, D.C. (0-referensedistrict)]
lgta	log of total assets
age	Age (measured in months or quarters)
State-Specific macroeconomic variables	
ur	Unemployment rate
chpi	Percentage change in personal income
chphi	Percentage change in house price index
chnphu	Change in new private housing units authorized by building permits

positive effect on a bank's survival. Higher ratios may signal banks to anticipate difficulties in recovering losses and thus this variable may positively impact incidence. Similarly, charge-offs on loan and lease loss recoveries provide a signal of problematic assets that increase the probability of insolvency and failure. Provision for loan and lease losses are based upon management's evaluation of loans and leases that the reporting bank intends to hold. Such a variable can expect to decrease the probability of distress and increase the probability of survival. We can also view this as a proxy to control for one of the several ways in which different banks pursue different risk strategies (Inanoglu et al. 2014). An often-used measure of credit risk is the gross charge-off rate (dollar gross charge-offs normalized by lending book assets). We control for risk-taking strategies in which banks may engage that differ from their role as a provider of intermediation services–the service we analyze–by including both of these risk measures as explanatory variables.

Two of the three management quality proxies that we include are constructed from the balance sheet items of the reporting banks. The first is the ratio of the full-time employees to average assets, which has an ambiguous sign in both the incidence and latency parts of our model. We conjecture, however, a negative sign on this variable as the FDIC may face constraints in seizing large banks with a large number of employees. The second is the intermediation ratio, which shows the ability of a bank to successfully transform deposits into loans and thus we expect its overall impact also to be negative. Finally, the third management quality proxy is managerial performance, which we estimate as part of our combined model. The level of banks' earnings as measured by the operating income and returns on assets and equity are also expected to have a negative effect on both the incidence and latency parts. From liquid assets we expect cash and core deposits to have negative

signs, while the direction of the effect of Jumbo time deposits is uncertain. Banks with relatively more market price sensitive liabilities and illiquid assets should be considered at a higher risk of failure *ex ante*.

14.5 Results and Predictive Accuracy

In Table 14.3, we report the results for Model I and Model II. Both models produce qualitatively similar results. The influential factors that were considered to have a strong effect on both sets of probabilities a priori turn out to have the correct sign and most are statistically significant in both models. Results indicate that there is a large marginal effect of the tier 1 capital ratio on the incidence probability. Other measures of earnings proxies and asset quality also have a material effect on this probability. In other words, well-capitalized banks with positive earnings and quality loans are less likely to appear on the FDIC watch list. In contrast, banks that already are on this list will increase their probability of failure if their capital ratio is insufficient, their ratio of nonperforming loans is high, and their earnings are negative and have a decreasing trend. The certificates of deposits and core deposits have the expected effect though not a statistically significant one. On the other hand, cash has a positive and significant effect. One explanation of this could be that, after controlling for profitability, banks that remain cash idle have a higher opportunity cost. It would only stand to reason for these banks to be costly and inefficient. Banks with a large number of full-time employees are shown to have less chances to fail. Banks that successfully transform deposits into vehicles of investment are considered potentially stronger, while others with more rate sensitive liabilities appear to be less promising.

The state-specific variables have the expected economic congruences which appear to be non-significant in the incidence part of the models. We would expect these variables to significantly affect the probability of incidence of banks in the states with higher unemployment rates, lower growth in personal income, limited construction permits, and falling housing prices, all of which would give cause for increased on-site inspections. Only two of the four geographical variables have a significant effect. Banks that are Federal Reserve System (FRS) members have a higher probability of failure than those that are not. This is associated with behavior consistent with moral hazard. Such banks have felt secure as members of the FRS and hence may have assumed higher risks than they would have had they not been FRS banks. The positive result of the FRS district code indicates that the probability of insolvency and failure is higher for banks in the Atlanta (6) district than for banks in the Boston (1) district, for example. Bank size is shown to have a negative and significant effect only in the incidence part of Model II, which could be interpreted that larger banks are less likely to be placed on the watch list and subsequently fail. Finally, the older and well-established banks appear to have lower failure probabilities than their younger counterparts.

Table 14.3 Parameter estimates obtained under Model I (SPMHM) and Model II (DTMHM)

Variable	Model I		Model II	
	Latency	Incidence	Latency	Incidence
Intercept		−2.5989 (2.8512)		4.9130 (2.9266)
lgta	0.0797 (0.0885)	0.0607 (0.1103)	0.0531 (0.0875)	−0.3320*** (0.1056)
age	−0.0004* (0.0002)	0.0004 (0.0003)	−0.0003 (0.0002)	0.0001 (0.0003)
tier 1	−48.417*** (3.0567)	−86.791*** (5.3156)	−47.060*** (3.0856)	−88.728*** (5.3516)
alll	−9.5829** (4.7047)	16.473** (7.9759)	−8.8615* (4.6962)	8.8671 (7.7594)
reln	4.4321*** (1.1801)	2.0116 (1.2748)	3.7811*** (1.1731)	3.9762*** (1.2796)
rnpl	7.2555*** (1.3348)	6.3838*** (2.1574)	6.1802*** (1.3433)	9.6447*** (2.1510)
roa	−6.1672 (5.1795)	−11.248** (5.5416)	−7.2727 (5.0983)	−8.8145 (6.1201)
roe	0.0003 (0.0003)	0.0002 (0.0013)	0.0003 (0.0004)	0.0003 (0.0017)
cd	1.0098 (0.8644)	1.6651 (1.0274)	1.2499 (0.8425)	0.8245 (1.0003)
coredep	−2.7654 (1.7546)	−1.2140 (2.0839)	−2.5466 (1.7496)	−3.1272 (2.0927)
coffs	0.2351*** (0.0804)	0.3168*** (0.1183)	0.2319*** (0.0848)	0.2703** (0.1243)
lrec	38.162** (18.672)	14.463 (56.448)	35.681* (21.726)	37.945 (42.003)
llp	−10.427** (4.9577)	−15.501** (6.5158)	−11.688** (4.8628)	−13.155** (6.6190)
fte	−0.8228 (1.0468)	−3.0004** (1.4396)	−0.8329 (1.0512)	−3.1287** (1.4021)
imr	−4.2141*** (1.0634)	−1.7238 (1.2254)	−3.7792*** (1.0603)	−4.4020*** (1.2016)
sens	2.3255*** (0.8386)	2.5869** (1.0320)	2.0025** (0.8403)	5.6444*** (1.0042)
cash	6.7983*** (2.0542)	6.7497** (3.2628)	6.9211*** (2.0472)	4.5465 (3.6333)
oi	−3.9670 (4.4353)	−6.1756 (6.6955)	−3.1670 (4.3948)	−4.9651 (6.6585)
ur	0.1198*** (0.0379)	0.0196 (0.0490)	0.0655* (0.0390)	0.0548 (0.0482)
chpi	−15.091* (8.1323)	−10.555 (9.7490)	−20.313** (8.0823)	−10.645 (10.017)
chhpi	−8.1375* (4.9453)	−3.1678 (5.8411)	−9.8817** (4.8428)	−5.4824 (5.8215)

(continued)

Table 14.3 (continued)

Variable	Model I		Model II	
	Latency	Incidence	Latency	Incidence
chnphu	−0.6570*** (0.2490)	0.0006 (0.0523)	−0.5246** (0.2417)	0.0047 (0.0581)
chtype	−0.2151 (0.5058)	0.4441 (0.6871)	0.0223 (0.5051)	−0.7143 (0.5943)
frsmb	0.4707*** (0.1797)	0.4018* (0.2352)	0.4617*** (0.1808)	0.3466 (0.2363)
ibf	1.1171 (0.7589)	1.4405 (0.8883)	1.2816* (0.7592)	−2.5959*** (0.7825)
frsdistrcode	0.2465*** (0.0329)	0.2615*** (0.0430)	0.2295*** (0.0325)	0.2457*** (0.0427)
LogL	1763.87		1714.92	
N	5968		38,571	

$p^* < 0.1$, $p^{**} < 0.05$, $p^{***} < 0.01$ (Robust standard errors in parentheses)

In Table 14.4, we present results for the continuous-time semiparametric and discrete-time MHM with the stochastic frontier specification. With few exemptions, the results are qualitatively similar to those reported in Table 14.3. Inefficiency has a positive effect on the incidence and failure probabilities. The effect is only significant on the latter probability and this is consistent with the view that bank performance is not the criterion for an on-site examination, but rather a factor affecting a bank's longer term viability. The distributional parameters are significant at the one-percent significance level. The descriptive statistics for the efficiency score obtained from Models III and IV as well as from the standard time-invariant random effects (RE) model for the sample of nonfailed and failed banks are summarized in Table 14.5. There is a small, but a statistically significant difference between the average efficiencies estimated form Models III and IV. This difference is not statistically significant for efficiencies derived from the RE model. Figure 14.1 depicts the distribution of inefficiencies obtained from the three models (Model III, Model IV and RE). It is worthwhile to note that the RE model reports certain surviving banks as extremely inefficient, while the most efficient banks are banks that failed. Based on these observations, we suspect that the two-step approach would yield the opposite sign on inefficiency component from what we would expect. The difference in average efficiencies from the single step estimation can be mainly attributed to the fact that distressed banks typically devote their efforts to overcome their financial difficulties and clean up their balance sheets. These impose additional costs on banks and worsen their already bad financial position.

In Figs. 14.2 and 14.3 we depict the survival profile of the average bank that failed during the 2008–2009 period for all four models. It can be seen from Fig. 14.2 that average failed banks in Model I are predicted to have a duration time of twenty two months. After controlling for inefficiencies, the time to failure drops

Table 14.4 Parameter estimates obtained under Model III (SPMHM + SF) and Model IV (DTMHM + SF)

Variable	Model III		Model IV	
	Latency	Incidence	Latency	Incidence
Intercept		−2.6934 (2.8039)		4.7694* (2.9201)
lgta	−0.0408 (0.0935)	−0.0087 (0.1243)	−0.0742 (0.0958)	0.3466*** (0.1117
age	−0.0004* (0.0002)	0.0004 (0.0003)	−0.0004* (0.0002)	0.0001 (0.0003)
tier 1	−48.647*** (3.0631)	−86.280*** (5.3068)	−48.452*** (3.0678)	−88.684*** (5.339)
alll	−8.5073* (4.6488)	17.003** (7.9738)	−8.8587* (4.6066)	8.8881 (7.7741)
reln	4.6588*** (1.1259)	2.1044* (1.2631)	4.4871*** (1.0878)	3.9288*** (1.2805
rnpl	6.9014*** (1.3206)	6.0653*** (2.1661)	6.7347*** (1.3210)	9.5835*** (2.1554
roa	−6.4175 (5.0868)	−11.451*** (5.5328)	−6.1672 (5.1423)	−8.8129 (6.1180)
roe	0.0002 (0.0004)	0.0001 (0.0013)	0.0002 (0.0003)	0.0002 (0.0017)
cd	0.8641 (0.8608)	1.5840 (1.0277)	0.7565 (0.8579)	0.7329 (1.0244)
coredep	−2.3913 (1.6224)	−1.0244 (2.0568)	−1.5432 (1.6367)	−2.9196 (2.1285)
coffs	0.2447*** (0.0801)	0.3232*** (0.1172)	0.2516*** (0.0798)	0.2720** (0.1243
lrec	37.309** (19.148)	14.661 (56.167)	37.219** (19.011)	38.569 (41.568)
llp	−11.175** (4.9577)	−15.784** (6.5199)	−11.654** (4.7671)	−13.211** (6.6345
fte	−2.1781** (1.0195)	−3.8780** (1.5932)	−2.8670*** (1.0298)	−3.3559** (1.5165
imr	−3.7660*** (0.9728)	−1.4553 (1.2128)	−3.2640*** (0.9832)	4.2466*** (1.2601
sens	2.2143*** (0.8294)	2.5264** (1.0309)	2.0894** (0.8282)	5.6036*** (1.0068
cash	7.4166*** (2.0368)	7.1461** (3.2558)	7.6605*** (2.0375)	4.6012 (3.6396)
oi	−3.9483 (4.3968)	−6.4722 (6.6946)	−4.1980 (4.3825)	−5.0126 (6.6601)
ur	0.1210*** (0.0377)	0.0234 (0.0491)	0.1208*** (0.0378)	0.0555 (0.0487)
chpi	−15.567** (8.1081)	−9.7061 (9.7811)	−15.551* (8.0642)	−10.639 (10.021)
chhpi	−8.1886* (4.9593)	−3.2387 (5.8526)	−8.1802** (4.9731)	−5.4864 (5.8139)

(continued)

Table 14.4 (continued)

Variable	Model III		Model IV	
	Latency	Incidence	Latency	Incidence
chnphu	−0.6300*** (0.2471)	−0.0001 (0.0531)	−0.6171*** (0.2456)	0.0046 (0.0578)
chtype	−0.1496 (0.5045)	0.4875 (0.6876)	−0.1293 (0.5028)	−0.7224 (0.5953)
frsmb	0.4960** (0.1801)	0.3994* (0.2349)	0.4977*** (0.1801)	0.3487 (0.2512)
ibf	1.1325 (0.7574)	1.4718* (0.8945)	1.1295* (0.7571)	−2.5923*** (0.7815)
frsdistrcode	0.2612*** (0.0332)	0.2725*** (0.0445)	0.2663*** (0.0334)	0.2469*** (0.0429
δ1		0.2062 (0.1577)		0.0343 (0.0828)
δ2	0.3058*** (0.0468)		0.4137*** (0.0750)	
σ	0.0552*** (0.0011)		0.0548*** (0.0011)	
Y	0.5173*** (0.0017)		0.5278*** (0.0015)	
LogL	67,701		66,360	
N	5968		38,571	

$p^* < 0.1$, $p^{**} < 0.05$, $p^{***} < 0.01$ (Robust standard errors in parentheses)

Table 14.5 Cost efficiencies results

	Mean	Standard deviation	Minimum	Maximum
Non failed banks				
Model III	0.6817	0.0691	0.3167	0.9705
Model IV	0.7295	0.1630	0.1992	0.9688
Random effects	0.6466	0.0662	0.4636	0.9650
Failed banks				
Model III	0.6721	0.1022	0.1499	0.8722
Model IV	0.6804	0.0824	0.1539	0.8488
Random effects	0.6408	0.0798	0.3845	0.8626

The top and bottom 5% of inefficiencies scores are trimmed to remove the effects of outliers

to twenty one months. Based on the Model II results, Fig. 14.3 demonstrates that a bank with the same characteristics as the representative failed bank will survive up to 7 quarters, after accounting for inefficiency.

It is also interesting to look at the survival profile of the most and the least efficient banks derived from Model III and Model IV. Figure 14.4 displays the survival profiles obtained from Model III. The least efficient bank with an efficiency score of 0.149% and is predicted to fail in eight months. This bank was closed by FDIC in the end of August of 2008. On the other hand, the most efficient bank with efficiency score of 0.971% has a survival probability of one throughout the sample

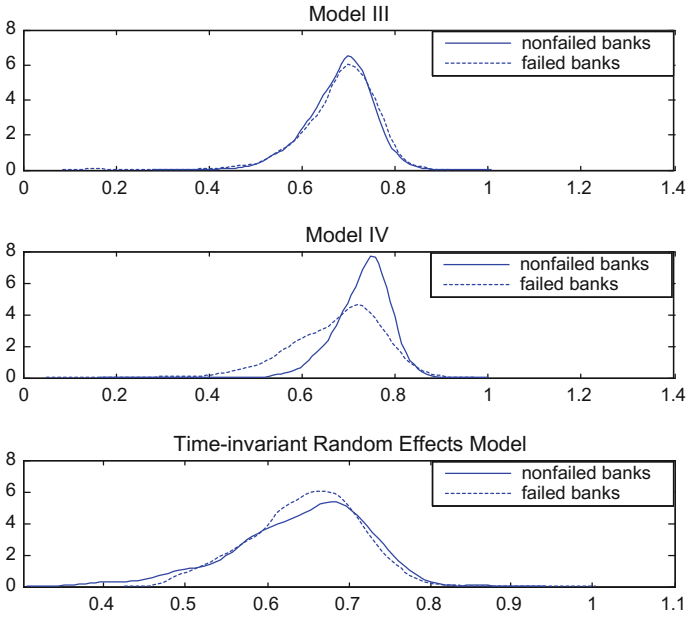


Fig. 14.1 Cost efficiency distributions

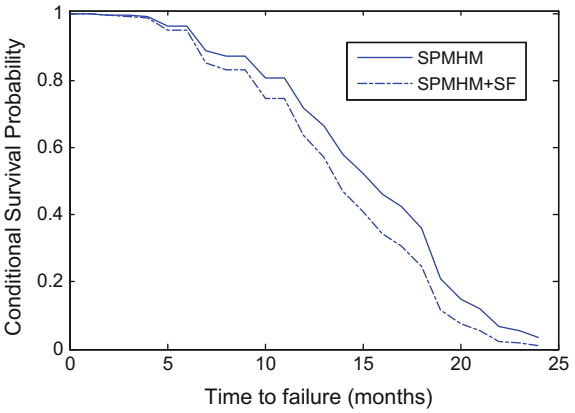


Fig. 14.2 SPMHM: Failed banks' average survival pofile

period. This is also illustrated in Fig. 14.5, where the least efficient bank with an efficiency score of 0.154% is predicted to fail by fifth quarter, using the Model IV results. This bank failed in the third week of April of 2009.[29] The most efficient bank with an efficiency score of 0.969 has an estimated survival probability that exceeds 95%.

[29]The least efficient bank is not the same in these two models. However, the most efficient bank is.

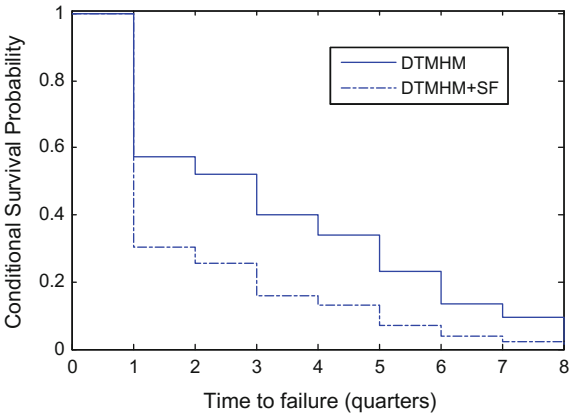


Fig. 14.3 DTMHM: Failed banks' average survival pofile

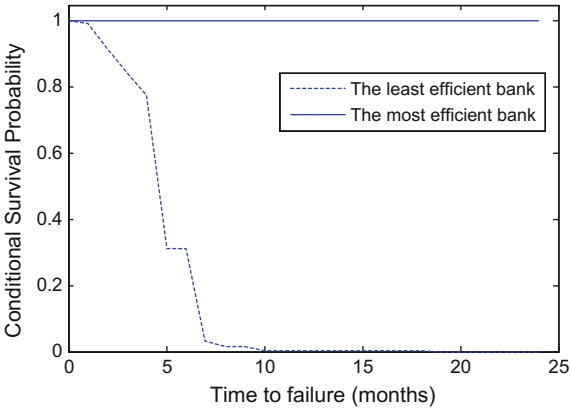


Fig. 14.4 Model IV: The most and the least efficient bank's survival profile

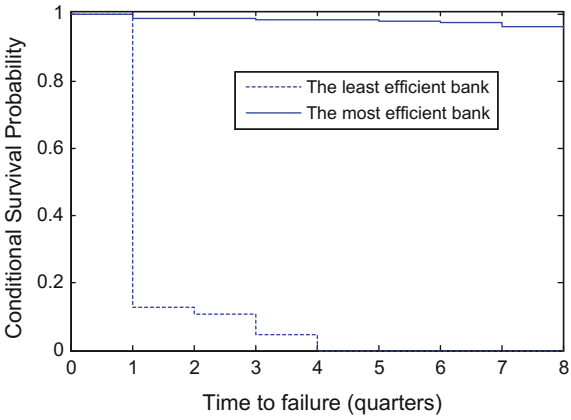


Fig. 14.5 Model III: The most and the least efficient bank's survival profile

We next examine our results by recasting our model estimates as early warning tools that can correctly classify failed and nonfailed banks within our sample used for estimation as well as in our hold-out 2010–2011 sample. The tests are based on two types of errors, similar to those that arise in any statistical hypothesis testing. These are type I and type II errors.[30] A type I error is defined as the error due to classifying a failed bank as a nonfailed bank, while a type II error arises from classifying a non-failed bank as a failed bank. There is a trade-off between these two type of errors and both are important from a public policy standpoint. Models with low type I error are more desirable, since timely identification of failed banks allows the regulator to undertake any prompt corrective action to ensure that the stability and the soundness of the financial system is not compromised. On the other hand, models with high type II error unnecessary will be flagging some banks as failures while they are not, and hence could waste the regulators' time and resources. However, it is oftentimes hard to interpret the costs of a type II error since various constraints faced by the FDIC could delay the resolution of an insolvent bank. Thompson (1992) attributes this to information, administrative, legal and political constraints, among others. Whalen (1991) notes that some type II error predictions actually represent failures that occur in the near future and hence should be considered as a success of the model rather than its failure.

In Table 14.6, we report the in-sample predictive accuracy for the four models based on type I, type II, and overall classification error. Overall classification error is a weighted average of type I and type II errors. In what follows we set the weights at 0.5 for both errors.[31] In our predictive accuracy analysis, each bank is characterized as a failure if its survival probability falls bellow a probability cutoff point, which we base on the sample average ratio of failed to nonfailed banks (0.021). The results in Table 14.6 indicate that the discrete-time specification yields a lower type I error than does the continuous-time specification. This is to be expected since the former incorporates multi-period observations for each bank and thus is more informative about a bank's financial health than the single-period cross-sectional observations. There is a significant drop in type I error in both specifications when the performance of a bank is added to the model as an additional factor. On the other hand type II error is increased in the discrete-time models and it is doubled when inefficiency is included. Based on the overall classification error, Model IV performs somewhat better than Model III, but it largely outperforms Models I and II.

Table 14.6 also presents the errors that judge the 2010–2011 out-of-sample classification accuracy of our models based on the estimates obtained using 2008–2009 data. The continuous-time models' errors are based on the estimated survival profiles of banks using the 2009 end-year data, while the discrete-time models' errors use the full 2010–2011 data. By comparing these errors with the 2008–2009

[30]See (Lane et al. 1986; Whalen 1991; and Thompson 1992) among others.

[31]Clearly this weighting scheme is arbitrary and alternative weighting schemes could be based on different risk preference assumptions, implicit and explicit costs of regulation, etc.

Table 14.6 Predictive accuracy results

	Model I	Model II	Model III	Model IV
2008–2009 in-sample classification				
Type I error	0.3840	0.2882	0.1123	0.0644
Type II error	0.0047	0.0051	0.0231	0.0476
Overall classification error	0.1937	0.1465	0.0581	0.0573
2010–2011 out-of-sample classification				
Type I error	0.2283	0.0292	0.1630	0.0292
Type II error	0.0049	0.0012	0.0062	0.0012
Overall classification error	0.1157	0.0152	0.0840	0.0152
2010–2011 in-sample classification				
Type I error	0.1988	0.0702	0.1404	0.0468
Type II error	0.0025	0.0012	0.0025	0.0012
Overall classification error	0.1007	0.0357	0.0715	0.0240

Overall classification error is a simple average of type I and type II errors

in-sample classification errors, we observe that there is a significant drop in type I error for all four models. This is may be due to the fact that the data used to estimate the banks' survival profiles are more informative than what was used to estimate the model parameters, which is reasonable given that the end of 2009 was considered the peek year of the 2007–2011 banking crisis. The inter-model comparison is the same as above with Model IV favored over the other models based on predictive accuracy. In addition, all four models predict the major (i.e., with total assets size over $1 billion) and the minor bank failures equally well, by reporting very low estimated type I errors. In fact, type I error is zero for all major in- and out-of-sample bank failures.[32]

In order to examine the sensitivity of the models' classification accuracy to the data period selection (high risk period versus low risk period), we also estimate the models' in-sample classification accuracy using 2010–2011 data.[33] The 2010–2011 in-sample classification errors are also summarized in Table 14.6. Comparing the 2010–2011 out-of-sample results to the 2010–2011 in-sample results, we observe that type I error is slightly decreased for the continuous-time models (by 0.0295 in Model I and by 0.0226 in Model III), but it is increased in the discrete-time models (by 0.041 in Model II and by 0.0176 in Model IV). More specifically, Model II fails to predict the failure of 12 out of 171 banks that failed in our 2010–2011 sample, while Model IV fails to predict the failure of 8 out of 171 failed banks during the same period. The corresponding 2010–2011 out-of-sample predictions failed to identify only 5 of such failures. Overall, the predictive power of our models appears to be quite robust across different estimation sub-periods within the current financial crisis. We note, however, that conditions that led to the 2007–2011 banking crisis may be substantially different from those of future

[32]Detailed survival profile series for each bank in our sample are available upon request.

[33]The parameter estimates from the 2010–2011 estimation are available upon request.

banking crises. In this case, not only the model estimates, but also the variables that are used to predict banking troubles and failures can significantly differ.

14.5.1 Endogenous Variables and Identification

Two potential complications naturally may arise in structural models like the ones presented in this paper: the presence of endogenous variables and issues of identification of the structural model. Testing for potential endogenous control variables from our variable list and identification of the casual effect of the efficiency component is the purpose of this subsection.

First, we note that some of the control variables from our list of covariates may be potentially treated as endogenous in the sense that these are under a bank's management control and potentially can be affected by the probability and timing of failure. In particular, there is the possibility that a bank that is placed on the FDIC's watch list may erroneously report (underestimate or overestimate) the amount of these variables in its Call Reports. Such variables may include the provision for loan and lease losses,[34] which involves subjective assessment by a bank's management, and the number of the full-time employees, which is subject to substantial variation during distressed times. Other variables, such as allowance for loan and lease loss, charge-offs on loans and leases and recoveries on allowance for loan and lease losses also can be treated as endogenous. However, we note that these are subject to stringent scrutiny by regulators and auditors who can recognize and measure the effectiveness and appropriateness of management's methodology for collectively and individually assessing these accounts in accordance with internationally accepted reporting standards. We, therefore, treat these variables as exogenous in our models and do not further test for their exogeneity.

Below we do test for the endogeneity of the provision for loan and lease losses and the number of the full-time employees. We use a nonparametric test based on Abrevaya, Hausman and Khan (2010). The test is carried out using the following steps:

- Step 1: Identify, select and validate instrumental variables for the potentially endogenous variables;
- Step 2: Project the potentially endogenous variables onto the column space of the instrumental and exogenous variables and obtain their fitted values;
- Step 3: Estimate the model[35] separately by using the potentially endogenous variables and instrumented endogenous variables and obtain the survival profiles under both cases (label these as S_end and S_iv, respectively)

[34]The provision for loan and lease loss is the amount required to establish a balance in the allowance for credit losses account, which management considers adequate to absorb all credit related losses in its loan portfolio.

[35]Note that for testing purposes only the time-varying model combined with efficiencies (i.e., Model IV) is used.

- Step 4: Use Kendall's tau rank correlation statistic to test for association/dependence of S_end and S_iv (i.e., test for the null that S_end and S_iv are not associated/dependent)
- Step 5: Reject the null hypothesis of endogeneity if the *p*-value of Kendall's tau statistic is below the desired confidence level.

For the provision for loan and lease losses/average loans and leases variable, the selected instruments are (i) one period lagged values of the provision for loan and lease losses/average loans and leases; (ii) one period lagged values of the non-performing loans/total loans; (iii) one period lagged values of the allowance for loan and lease loss/average loans and leases; and (iv) one period lagged values of the recoveries on allowance for loan and lease losses/average loans and leases. The estimated Kendall's tau statistic is 0.9293 (with *p*-value = 0) and thus we reject the null hypothesis that the provision for loan and lease losses variable is endogenous in our estimation sample. Similarly, for the number of full-time equivalent employees/average assets variable, the selected instruments are (i) current period overhead expense; (ii) one period lagged values of the overhead expense; (iii) current period ratio of non-interest expense/total assets; (iv) one period lagged values of the ratio of non-interest expense/total assets; (v) current period ratio of non-interest expense/interest expense; and (vi) one period lagged values of the ratio of non-interest expense/interest expense. The estimated Kendall's tau statistic is 0.9723 (with *p*-value = 0) and we thus reject the null hypothesis that the number of the full-time equivalent employees variable is endogenous in our estimation sample. Joint testing yields Kendall's tau statistic of 0.9115 (with p-value = 0), thus leading to the same conclusion that both of these variables are not endogenous in our sample.

To corroborate the testing results above, we also test for the endogeneity of the provision for loan and lease losses and the number of the full-time employees by considering only the incidence part of the model. The rationale for using this alternative testing approach is that one might consider that these variables would be affected primarily by the incidence probability, as a bank's management could potentially manipulate these accounts to avoid being placed on the FDIC's watch list in the first place. The testing results are based on the Wald statistic on the hypothesis is exogeneity of the potential endogenous variables.[36] The Wald statistic is 1.74 (with *p*-value = 0.1866) for the provision for loan and lease losses/average loans and leases and 3.06 (with *p*-value = 0.0804) for the number of the full-time employees/average assets. These estimated test statistics are not significant at the 5% confidence level and generally corroborate the findings using the alternative null hypothesis.

The identification of the casual effect of the efficiency term, on the other hand, is performed by testing for the over-identifying restrictions using the testing approach outlined above. Due to the fact that the efficiency term is latent (unobserved) in our models, we use the efficiency scores obtained from the random effects (RE) model

[36]This testing is carried out by using STATA's ivprobit command.

as a proxy for the combined model's efficiencies. We identify the one period lagged values of the return on assets, the one period lagged values of the return on equity, the one period lagged value of the intermediation ratio (total loans/total deposits), the ratio of non-interest expense/interest expense, and the one period lagged values of the ratio of non-interest expense/interest expense as instrumental variables for the estimated efficiency scores. The resulting Kendall's tau statistic is estimated as 0.7970 (with p-value = 0); thus, rejecting the null hypothesis that the casual effects are not identified in our estimation sample.

14.6 Concluding Remarks

Massive banking failures during the financial turmoil of the Great Recession has resulted in enormous financial losses and costs to the U.S. economy, not only in terms of bailouts by regulatory authorities in their attempt to restore liquidity and stabilize the financial sector, but also in terms of lost jobs in banking and other sectors of economy, failed businesses, and ultimately slow growth of the economy as a whole. The design of early warning models that accurately predict the failures and their timing is of crucial importance in order to ensure the safety and the soundness of the financial system. Early warning models that can be used as off-site examination tools are useful for at least three reasons. They can help direct and efficiently allocate the limited resources and time for on-site examination so that banks in immediate help are examined first. Early warning models are less costly than on-site visits made by supervisors to institutions considered at risk and can be performed with high frequency to examine the financial condition of the same bank. Finally, early warning models can predict failures at a reasonable length of time prior to the marked deterioration of a bank's condition and allow supervisors to undertake any prompt corrective action that will have minimal cost to the taxpayer.

In this paper we have considered early warning models that attempt to explain recent failures in the U.S. commercial banking sector. We employed a duration analysis model combined with a static logit model to determine troubled banks which subsequently fail or survive. Both continuous and discrete time versions of the mixed model were specified and estimated. These effectively translated the bank-specific characteristics, state-related macroeconomic variables, and geographical and market structure variables into measures of risk. Capital adequacy and nonperforming loans were found to play a pivotal role in determining and closing insolvent institutions. State-specific variables appeared to significantly affect the probability of failure but not insolvency. The discrete-time model outperformed the continuous-time model as it is able to incorporate time-varying covariates, which contain more and richer information. We also found that managerial efficiency does not significantly affect the probability of a bank being troubled but plays an important role in their longer term survival. Inclusion of the efficiency measure led to improved prediction in both models.

Acknowledgements We would like to thank Mahmoud El-Gamal, John Bryant, Natalia Sizova, Qi Li, and Bryan Brown for their insightful comments and suggestions. We also thank seminar participants at the Economic Workshops and Seminars at Rice, Texas A&M University and the University of Queensland for their helpful suggestions and criticisms. We would like to thank Robert Adams at the Board of Governors of the Federal Reserve System for his valuable help in data construction. The views expressed in this paper are solely the responsibility of the first author and are independent from views expressed by Ernst & Young, LLP. Any errors are our own.

Appendix

In this appendix we show the derivation of the sample likelihood function given in expression (14.3). For this purpose we first note that at time t, bank i can fall into four mutually exclusive states of nature:

$$States = \begin{cases} h_i = 1,\ d_i = 1\ (\text{Problem \& Failed}) & \text{with prob. } F_e(x_i'\boldsymbol{\beta})\lambda_i^p(t;w_i)S_i^p(t;w_i) \\ h_i = 0,\ d_i = 1\ (\text{Sound \& Failed}) & \text{with prob. } \left[1 - F_e(x_i'\boldsymbol{\beta})\right]\lambda^s(t;w_i)S_i^s(t;w_i) \\ h_i = 1,\ d_i = 0\ (\text{Problem \& Censored}) & \text{with prob. } F_e(x_i'\boldsymbol{\beta})S_i^p(t;w_i) \\ h_i = 0,\ d_i = 0\ (\text{Sound \& Censored}) & \text{with prob. } \left[1 - F_e(x_i'\boldsymbol{\beta})\right]S_i^s(t;w_i) \end{cases}$$

Then

$$\begin{aligned} L(\boldsymbol{\theta};x,w,d) &= \prod_{i=1}^{n} L_i(\boldsymbol{\theta};x,w,d) \\ &= \prod_{i=1}^{n} \left\{ \left[F_e(x_i'\boldsymbol{\beta})\lambda_i^p(t;w_i)S_i^p(t;w_i)\right]^{h_i} \left(\left[1 - F_e(x_i'\boldsymbol{\beta})\right]\lambda_i^s(t;w_i)S_i^s(t;w_i)^{1-h_i} \right) \right\}^{d_i} \\ &\quad \times \left\{ \left[F_e(x_i'\boldsymbol{\beta})S_i^p(t;w_i)\right] \left({}^{h_i}\left[1 - F_e(x_i'\boldsymbol{\beta})\right]S_i^s(t;w_i)\right)^{1-h_i} \right\}^{1-d_i} \\ &= \prod_{i=1}^{n} F_e(x_i'\boldsymbol{\beta})^{h_i}\left[1 - F_e(x_i'\boldsymbol{\beta})\right]^{(1-h_i)}\left[\lambda_i^p(t;w_i)\right]^{d_i h} \\ &\quad \times \left[\lambda_i^s(t;w_i)\right]^{d_i(1-h_i)}\left[S_i^p(t;w_i)\right]^{h_i}\left[S_i^s(t;w_i)\right]^{1-h_i} \end{aligned}$$

By assumption, $\lambda_i^s(t;w_i) = 0$, if and only if, $h = 0$ and $d_i = 0$ (i.e., a bank is healthy and is not observed failing). Similarly $S_i^s(t;w_i) = 1$ if and only if $h_i = 0$ (i.e., a bank is healthy). The final sample likelihood function is then given by

$$L(\theta;x,w,d) = \prod_{i=1}^{n} F_e(x_i'\boldsymbol{\beta})^{h_i}\left[1 - F_e(x_i'\boldsymbol{\beta})\right]^{(1-h_i)}\left[\lambda_i^p(t;w_i)\right]^{d_i h_i}\left[S_i^p(t;w_i)\right]^{h_i}$$

which implies that the completely healthy banks contribute to the likelihood function only through their probability being troubled.

References

Abrevaya J, Hausman JA, Khan S (2010) Testing for causal effects in a generalized regression model with endogenous regressors. Econometrica 78(6):2043–2061

Aigner DJ, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier models. J Econometrics 6:21–37

Altman EI (1968) Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. J Finance 23:589–609

Baltensperger E (1980) Alternative approaches to the theory of the banking firm. J Monetary Econ 6:1–37

Barr RS, Siems TF (1994) Predicting bank failure using DEA to quantify management quality. Financial industry studies working paper 94–1, Federal Reserve Bank of Dallas

Battese GE, Cora GS (1977) Estimation of a production frontier model: with application to the pastoral zone of eastern Australia. Aust J Agric Econ 21:169–179

Bover O, Arellano M, Bentolila S (2002) Unemployment duration, benefit duration and the business cycle. Econ J 112:223–265

Charnes A, Cooper WW, Rhodes EL (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444

Cole RA, Gunther JW (1995) Separating the likelihood and timing of bank failure. J Bank Finance 19:1073–1089

Cole RA, Gunther JW (1998) Predicting bank failures: a comparison of on- and off-site monitoring systems. J Fin Serv Res 13:103–117

Cole RA, Wu Q (2009) Predicting bank failures using a simple dynamic hazard model. FDIC working paper, Washington, DC

Cole RA, Wu Q (2011) Is hazard or probit more accurate in predicting financial distress? Evidence from U.S. bank failures. MPRA Paper No. 29182, Munich, Germany

Cornwell C, Schmidt P, Sickles RC (1990) Production frontiers with cross-sectional and time series variation in efficiency levels. J Econometrics 46:185–200

Cox DR (1972) Regression models and life-tables (with discussion). J Roy Stat Soc B 34:187–220

Cox DR, Oakes D (1984) Analysis of survival data. Chapman & Hall, New York

Deakin E (1972) A discriminant analysis of predictors of business failure. J Account Res Spring:167–179

Debreu G (1951) The coefficient of resource utilisation. Econometrica 19:273–292

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm (with discussion). J Roy Stat Soc B 39:1–38

DeYoung R (1999) Birth, growth, and life or death of newly chartered banks. Econ Perspect 23:18–35

DeYoung R (2003) The Failure of new entrants in commercial banking markets: a split-population duration analysis. Rev Fin Econ 12:7–33

Farewell VT (1977) A model for a binary variable with time-censored observations. Biometrika 64:43–46

Farewell VT (1982) The use of mixture models for the analysis of survival data with long-term survivor. Biometrics 38:1041–1046

Farrell M (1957) The measurement of productive efficiency. J Roy Stat Soc A Gen 120:253–281

Gonzalez-Hermosillo B, Pazarbasioglu C, Billings R (1997) Determinants of banking system fragility. IMF Staff Papers 44(3)

Greene WH (2008) The econometric approach to efficiency analysis, chapter 2. In: Fried HO, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency: techniques and applications. Oxford University Press, New York

Halling M, Hayden E (2006) Bank failure prediction: a two-step survival time approach. In: Proceedings of the international statistical institute's 56th session

Hausman JA, Taylor WE (1981) Panel data and unobservable individual effects. Econometrica 49:1377–1398

Inanoglu H, Jacobs M, Liu R, Sickles RC (2014) Analyzing bank efficiency: are "too-big-to-fail" banks efficient? Rice University, Mimeo

Jondrow J, Lovell CAK, Materov IS, Schmidt P (1982) On the estimation of technical inefficiency in the stochastic frontier production function model. J Econometrics 19:233–238

Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data, 2nd edn. Wiley, New York

Kaparakis EI, Miller SM, Noulas AG (1994) Short-run cost inefficiency of commercial banks: a flexible stochastic frontier approach. J Money Credit Bank 26:875–893

Kasa K, Spiegel MM (2008) The role of relative performance in bank closure decisions. Econ Rev, Federal Reserve Bank of San Francisco

Klein JP, Moeschberger ML (2003) Survival analysis. Techniques for censored and truncated data, 2nd edn. Springer, New York

Kuk AYC, Chen CH (1992) A mixture model combining logistic regression with proportional hazards regression. Biometrika 79:531–541

Kumbhakar S, Lovell CAK (2000) Stochastic frontier analysis. Cambridge University Press, Cambridge, MA

Lancaster T (1990) The econometric analysis of transition data. Cambridge University Press, Cambridge, MA

Lane W, Looney S, Wansley J (1986) An application of the Cox proportional hazards model to bank failure. J Bank Finance 10:511–531

Martin D (1977) Early warning of bank failure: a logit regression approach. J Bank Finance 1:249–276

McLachlan G, Krishnan T (1996) The EM algorithm and extensions. Wiley, New York

Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb-Douglas production functions with composed error. Int Econ Rev 18:435–444

Meyer PA, Pifer HW (1970) Prediction of bank failures. J Finance 25:853–868

Pitt M, Lee L (1981) The measurement and sources of technical inefficiency in the Indonesian weaving industry. J Dev Econ 9:43–64

Sealey S, Lindley JT (1977) Inputs, outputs, and a theory of production and cost at depository financial institutions. J Finance 32:1251–1266

Shumway T (2001) Forecasting bankruptcy more accurately: a simple hazard model. J Bus 74:101–124

Sickles RC, Taubman P (1986) An analysis of the health and retirement status of the elderly. Econometrica 54:1339–1356

Sy L, Taylor J (2000) Estimation in a Cox proportional hazards cure model. Biometrics 56:227–236

Thompson JB (1992) Modeling the regulator's closure option: a two-step logit regression approach. J Fin Serv Res 6:5–23

Topaloglu Z, Yildirim Y (2009) Bankruptcy prediction. Working paper, Graduate Center, CUNY, Department of Economics

Torna G (2010) Understanding commercial bank failures in the modern banking era. http://www.fma.org/NY/Papers/ModernBanking-GTORNA.pdf

Tsionas EG, Papadogonas TA (2006) Firm exit and technical inefficiency. Empirical Econ 31:535–548

Whalen G (1991) A proportional hazards model of bank failure: An examination of its usefulness as an early warning model tool. Econ Rev, Federal Reserve Bank of Cleveland, 21–31

Wheelock D, Wilson P (1995) Explaining bank failures: deposit insurance, regulation, and efficiency. Rev Econ Stat 77:689–700

Wheelock D, Wilson P (2000) Why do banks disappear? the determinants of U.S. bank failures and acquisitions. Rev Econ Stat 82:127–138

Yildirim Y (2008) Estimating default probabilities of cmbs with clustering and heavy censoring. J Real Estate Finance Econ 37:93–111

Chapter 15
A Decomposition of the Energy Intensity Change in Spanish Manufacturing

Pablo Arocena, Antonio G. Gómez-Plana and Sofía Peña

Abstract The excessive consumption of energy has become a major economic and environmental concern in many countries over the last two decades. A country's energy performance is typically proxied by the rate of aggregate energy intensity, calculated as the ratio of energy consumed to GDP. The index number decomposition analysis is the usual approach to analyze the changes in a country's aggregate energy intensity. In this paper we analyze the energy intensity change as well as the energy efficiency change by combining the index decomposition analysis approach with non-parametric frontier efficiency methods. We apply this framework to decompose and analyze the sources of the change observed in the energy intensity of Spanish manufacturing industries during the period 1999–2007.

Keywords Energy intensity · Energy efficiency · Technical change · Index decomposition analysis · Frontier methods

15.1 Introduction

The efficient use of energy and the promotion of energy savings have come to occupy a prominent place in the economic and environmental agenda of many countries. It has received particular attention in the European Union, where a considerable number of Directives and other legislative initiatives have been passed in the last two decades (e.g. EC 2006, 2012). Today, energy efficiency constitutes

P. Arocena (✉) · A.G. Gómez-Plana · S. Peña
Institute for Advanced Research in Business and Economics (INARBE),
Departamento de Gestión de Empresas, Universidad Pública de Navarra,
Campus de Arrosadia, 31006 Pamplona, Spain
e-mail: pablo@unavarra.es

A.G. Gómez-Plana
e-mail: agomezgp@unavarra.es

S. Peña
e-mail: sofia.pena@unavarra.es

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_15

one of the cornerstones of the European Union's 2020 strategy (EC 2010). The latest increase of energy prices and the increasing impact of energy costs on industrial competitiveness have further encouraged the need of generating output with less consumption of energy (IEA 2013; Deloitte 2010).

An economy's energy efficiency is typically proxied by the rate of the aggregate energy intensity, calculated as the ratio of energy consumed to GDP. Thus, the evolution of the energy intensity is seen as a direct indicator of the relationship between economic growth and energy use, and specifically, to identify whether there is a decoupling of energy consumption from economic growth. Note however that a decrease in energy intensity is not a synonym of energy savings or of lower energy consumption in absolute terms. A decrease in energy intensity may also occur if energy consumption grows at a lower rate than GDP, which is known as relative decoupling.

Hence, it is important to determine the factors that influence the evolution of the energy intensity. To that end, energy researchers have developed a number of index decomposition methodologies in the last decades. In essence, the most widespread way of decomposing the change of the energy intensity index with this approach allows decomposing the change in the aggregate energy intensity into two types of components: the structural effect (sometimes called product mix or compositional effect), and the sectoral energy intensity effect (often called intrasectoral energy intensity or efficiency effect).

The traditional index decomposition approach has however a fairly limited analytic power to assess the effect of a number of factors that are critical to understand the variation of energy productivity rates within an industry, such as the improvement in the technical energy efficiency or the reduction of waste in the use of energy, the technical progress, the change in the degree of vertical integration and capital-labor ratio, the change in the scale of operations, and the variation in the spatial arrangement of production.

In this chapter we provide a decomposition analysis of the energy intensity change that combines frontier efficiency methods with the conventional index decomposition approach. The proposed approach allows the identification of a more comprehensive set of factors that explain the observed variation in energy intensity. Furthermore, it addresses the analysis of energy efficiency as an integral part of energy intensity. We apply this framework to identify the sources of the variation of energy intensity in Spanish manufacturing between 1999 and 2007.

The rest of the paper is organized as follows. Section 15.2 reviews the relevant literature on energy intensity. Section 15.3 develops the decomposition of the energy intensity change. Section 15.4 describes the methods employed to implement the decomposition. Section 15.5 presents the data and variables employed in the analysis. The results are discussed in Sect. 15.6. Conclusions and final remarks are commented in Sect. 15.7.

15.2 Energy Intensity and Energy Efficiency

As stated above, the energy-to-GDP ratio, referred to as energy intensity, is the most popular measure used in energy efficiency studies. Actually, the change of the energy intensity ratio is not strictly a measure of the change of energy efficiency, but the change in the reciprocal of the energy productivity ratio. Such distinction will be made clear later.

In any case, the index decomposition analysis (IDA) is the usual approach to quantify the underlying factors that contribute to changes in energy intensity, energy consumption, and related CO_2 emissions over time.[1] Since the late 1970s, a variety of index decomposition methods have been developed in the energy and environmental fields. The earliest studies were based on Laspeyres, Paasche, Marschall-Edgeworth, and Fisher ideal indexes. Boyd et al. (1988) pioneered the index decomposition based on the Divisia index, and introduced the so-called arithmetic mean Divisia index method. Ang and Zhang (2000) and Ang (1995, 2004a) provide comprehensive surveys of this earlier literature. However all these index approaches have the drawback of leaving a residual term i.e. the product (or the sum) of the estimated factors is not exactly equal to the observed change in the aggregate, which complicates the interpretation of the results. Moreover, they are unable to handle zero values in the data set.

The logarithmic mean Divisia index (LMDI) method was introduced by Ang and Choi (1997), and since then has become by far the most popular IDA approach due to its superior properties and its ease in practical implementation. As demonstrated in various papers (Ang and Zhang 2000; Ang and Liu 2001; Ang 2004b), the LMDI method jointly satisfies the factor reversal test and the time reversal test, it is robust to zero and negative values, and is perfect in decomposition (i.e. it ensures null residual terms). Further, the LMDI decomposition has both additive and multiplicative formulations (see Ang and Zhang 2000; Ang 2004b, 2015 for detailed analysis on alternative LMDI models).

To illustrate the LMDI method let us define the aggregate energy intensity of one country in period t as the ratio between the energy consumed (E) and the output (Y) obtained in year t, i.e.

$$I_t = \frac{E_t}{Y_t} \tag{15.1}$$

The aggregate energy intensity can be expressed as a summation of the sectoral data

[1] An alternative approach is the structural decomposition approach (SDA), which uses the input–output table as a basis for decomposition. Reviews of SDA can be found in Su and Ang (2012), Hoekstra and van den Bergh (2003).

$$\mathrm{I_t} = \frac{E_t}{Y_t} = \sum_{\mathrm{i}} \frac{\mathrm{E_{i,t}}}{\mathrm{Y_{i,t}}} \frac{\mathrm{Y_{i,t}}}{\mathrm{Y_t}} = \sum_{\mathrm{i}} \mathrm{I_{i,t}} \mathrm{S_{i,t}} \tag{15.2}$$

where E_t is the total energy consumption; $E_{i,t}$ is the energy consumption in sector i; Y_t is the aggregate output; $Y_{i,t}$ is the output of sector i; $I_{i,t}$ is the energy intensity of sector i and $S_{i,t} = Y_{i,t}/Y_t$ is the production share of sector i.

The change in the aggregate energy intensity between period 0 and 1 can be expressed as

$$dI = \frac{I_1}{I_0} \tag{15.3}$$

We apply the multiplicative LMDI-II model (Ang and Choi 1997; Ang and Liu 2001) to decompose the aggregate energy intensity[2]:

$$dI = \frac{I_1}{I_0} = \left[\exp\left(\sum_i w_i \ln\left(\frac{I_{i,1}}{I_{i,0}} \right) \right) \right] \cdot \left[\exp\left(\sum_i w_i \ln\left(\frac{S_{i,1}}{S_{i,0}} \right) \right) \right] \tag{15.4}$$

where

$$w_i = \frac{L\left(\frac{E_{i,1}}{E_1}, \frac{E_{i,0}}{E_0} \right)}{\sum_i L\left(\frac{E_{i,1}}{E_1}, \frac{E_{i,0}}{E_0} \right)} \tag{15.5}$$

In (15.5) $E_{i,t}/E_t$ is the share of sector i in the aggregate energy consumption in period t, and L is the logarithmic mean function introduced by Vartia (1976) and Sato (1976), which is defined as[3]

$$L\left(\frac{E_{i,1}}{E_1}, \frac{E_{i,0}}{E_0} \right) = \frac{\frac{E_{i,1}}{E_1} - \frac{E_{i,0}}{E_0}}{\ln \frac{E_{i,1}}{E_1} - \ln \frac{E_{i,0}}{E_0}} \tag{15.6}$$

The first component in (15.4) is the intensity effect, and measures the impact associated with changes in the energy intensity of individual sectors. The second component in (15.4) is the so-called structural effect, which accounts for the impact of the change in the sectoral composition of the economy, i.e. the variation in the share of each sector in total GDP.

[2]Ang (2015) argues that the multiplicative model is the preferred model for decomposing intensity indicators, while the additive composition analysis procedure is more suited when used in conjunction with a quantity indicator. In any case, there exists a direct relationship between the additive and multiplicative decompositions (Ang 2004b).

[3]The use of the logarithmic mean is more widespread than in the energy efficiency decomposition literature. Thus, its use in the analysis of price and quantity indexes is discussed in detail by Balk (2008), while Balk (2010) makes use of it in measuring productivity change.

The LMDI method has been widely used to decompose changes in energy intensity, energy consumption and energy-related carbon emissions in many countries (see e.g. Mulder and Groot 2012; Fernández et al. 2013, 2014; Voigt et al. 2014 for recent applications). In the case of Spain a number of studies apply LMDI methods to analyze the energy intensity change of the whole country (Cansino et al. 2015; Fernández-González et al. 2003; Mendiluce 2007; Marrero and Ramos-Real 2008; Mendiluce et al. 2010), and that of specific regions (e.g. Ansuategui and Arto 2004; Colinet and Collado 2015).

The efficient consumption of energy has been equally analyzed from the literature on efficiency and productivity measurement from a somehow different perspective. Basically, the measurement of efficiency is based on the idea of comparing the actual performance of an economic unit with respect to the optimum performance that technology allows. This technological boundary is not however directly observable, so it must be empirically estimated from the data. Therefore, the efficiency of a company is determined by comparing their performance with that of the best observed performers, which define the efficient frontier.

Filippini and Hunt (2015) relate this literature, which is firmly based on the economic foundations of production, with the concept of energy efficiency. There are many examples of energy efficiency studies that use the two dominant approaches in the field of efficiency measurement: the parametric Stochastic Frontier Analysis (SFA) and the non-parametric Data Envelopment Analysis (DEA). For instance, Filippini and Hunt (2011, 2012) and Orea et al. (2015) investigate the energy efficiency in various countries with stochastic frontier analysis, while Zhou and Ang (2008), and Zhou et al. (2008) provide examples of measuring the energy efficiency by means of linear programming techniques. In the next section, we combine the LMDI decomposition referred to above with a non-parametric frontier efficiency approach.

15.3 Methodology

15.3.1 Decomposing Firm's Energy Intensity Change

Let us first define the energy intensity of firm[4] j in year t as the ratio between the energy that consumes (E_j) and the output (Y_j) obtained in year t, i.e.

$$I_{j,t} = \frac{E_{j,t}}{Y_{j,t}} \tag{15.7}$$

[4]Here, we refer to the 'firm' as any producer or economic unit. The economic unit can equally refers to a region, as we do in our empirical application.

The observed change in the energy intensity of firm j between period 0 and 1, can then be expressed as

$$dI_j = \frac{I_{j,1}}{I_{j,0}} = \frac{\frac{E_{j,1}}{Y_{j,1}}}{\frac{E_{j,0}}{Y_{j,0}}} \tag{15.8}$$

We decompose the change in energy intensity in (15.8) as the product of three elements

$$dI_j = \left[\frac{E_1}{\frac{E_1^*(y_1)}{E_0}{E_0^*(y_0)}}\right] \cdot \left[\frac{E_1^*(y_0)}{E_0^*(y_0)}\right] \cdot \left[\frac{\frac{E_1^*(y_1)}{y_1}}{\frac{E_1^*(y_0)}{y_0}}\right] = EECH_j \cdot TCH_j \cdot SCH_j$$
$$= \textit{Energy efficiency change} \cdot \textit{Technical change effect} \cdot \textit{Scale change effect} \tag{15.9}$$

The first component in brackets in (15.9) is the *Energy Efficiency Change* (*EECH*). This element is defined as the quotient of two ratios. The numerator represents the firm's energy efficiency in period 1, measured as the ratio of the observed energy consumption in period 1 (E_1) and the minimum (efficient) level of energy required to produce the observed output level in period 1, $E_1^*(y_1)$. A firm is energy efficient if this ratio is equal to one, whereas a value greater than 1 indicates an excessive consumption of energy in producing the current output level.

Similarly, the denominator captures the energy efficiency relative to period 0. Therefore, a value of $EECH_j$ lower (greater) than unity indicates that energy efficiency of firm j has increased (decreased) between year 0 and 1, and thereby has contributed to reduce (increase) the observed energy intensity rate of firm j.

The second component in (15.9) represents the *Technical Change effect* (*TCH*), measured at the output level of period 0. This term quantifies the variation in the energy intensity driven by the shift in the technology between period 0 and period 1. Thus, it compares the minimum amount of energy required to produce the output level y_0 in period 1, with the minimum quantity of energy that was needed in period 0 to produce the same output level. Therefore, a value of TCH_j lower (greater) than one indicates that technical progress (regress) has occurred between the two time periods, contributing to reduce the firm's energy intensity.

Finally, the third component in (15.9) is the *Scale Change effect* (SCH). This term accounts for the impact on the variation of energy intensity resulting from a change in the scale of operations, taking the technology of period 1 as a reference. Thus, it is defined as the ratio between the minimum quantity of energy per unit of output needed to produce y_1 in period 1, and the minimum energy quantity per unit of output needed to produce y_0 in the same period 1.

Figure 15.1 illustrates our decomposition in a single-(energy) input single-output case. The picture represents two production technologies prevailing in two different time periods. Particularly, the curve F^t represents the boundary (or frontier) of the production technology of period t. Thus, production frontier F^0

represents the minimum input that is required in period 0 to produce any given level of output or, alternatively, the maximum output that can be obtained in period 0 from any given input quantity.

The figure shows the observed output level and the energy consumption of firm A in period 0 (A^0) and period 1(A^1). Specifically, it reflects a situation where a decrease in the firm's energy intensity rate has occurred between period 0 and period 1 ($E_1/y_1 < E_0/y_0$). In this example, it is easy to see that part of the reduction in the intensity rate is due to the improvement of energy efficiency. The energy efficiency of firm A has improved because the observed energy consumption of firm A in period 1 (E_1) is closer to the efficient energy quantity E_1^*, denoted by point P in its contemporaneous production frontier (F^1), than what it was in period 0 (i.e. A^0 is relatively farther from the point C in F^0), and therefore $EECH < 1$.

Secondly, the upward shift of the production frontier reveals that a technological progress has occurred between the two time periods. Consequently the production of the output quantity y_0 requires a lower amount of energy in period 1 than the quantity that was needed in the previous period, i.e. $E_1^*(y_0) < E_0^*(y_0)$, and therefore $TCH = <1$. In Fig. 15.1, the energy savings due to the technical progress are represented by the horizontal distance between points D and C.

Finally, the impact of the change in the scale of operations on the variation of the energy intensity is reflected by the movement along the production frontier F^1 from point D $\left(y_0, E_1^*(y_0)\right)$ to point P $\left((y_1, E_1^*\right)$, which results in SECH < 1.

The Energy Efficiency Change (EECH) component in (15.9) can be further decomposed into two terms:

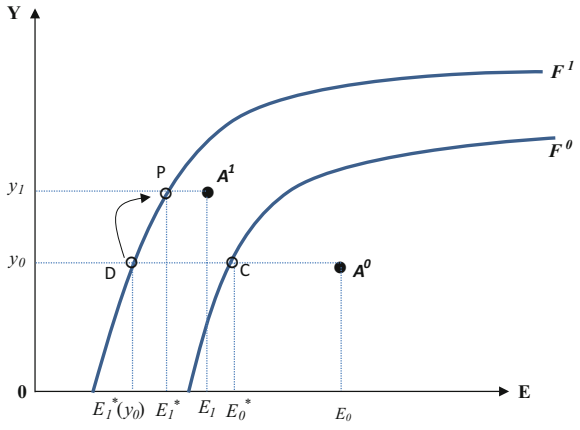


Fig. 15.1 Decomposing the change in energy intensity

$$EECH = \frac{\frac{E_1}{E_1^*(y_1)}}{\frac{E_0}{E_0^*(y_0)}} = \left[\frac{\frac{E_1}{E_1'(y_1)}}{\frac{E_0}{E_0'(y_0)}}\right] \cdot \left[\frac{\frac{E_1'(y_1)}{E_1^*(y_1)}}{\frac{E_0'(y_0)}{E_0^*(y_0)}}\right] \tag{15.10}$$
$$= Technical\ efficiency\ change \cdot Input\ mix\ change$$

Therefore, the full decomposition of the energy intensity change of firm j can be formulated as

$$dI_j = \left[\frac{\frac{E_1}{E_1'(y_1)}}{\frac{E_0}{E_0'(y_0)}}\right] \cdot \left[\frac{\frac{E_1'(y_1)}{E_1^*(y_1)}}{\frac{E_0'(y_0)}{E_0^*(y_0)}}\right] \cdot \left[\frac{E_1^*(y_0)}{E_0^*(y_0)}\right] \cdot \left[\frac{\frac{E_1^*(y_1)}{y_1}}{\frac{E_1^*(y_0)}{y_0}}\right] \tag{15.11}$$
$$= TECH_j \cdot IMCH_j \cdot TCH_j \cdot SCH_j$$

As Filippini and Hunt (2015) observe, there is not a unique and generally accepted definition of energy efficiency. Thus, a possible measure is the Farrell's radial input measure of technical efficiency (Farrell 1957), in which the improvement of the level of efficiency requires a proportional reduction in both energy and the other inputs. However, we are interested in measuring the specific efficiency in the use of energy, and to that end in (15.10) we have introduced two different energy efficient benchmark quantities. To explain the differences between them, let us consider a KLEM model, which defines output as a function of capital (K), labor (L), energy (E) and other intermediate inputs (M). Other intermediate inputs (M) include materials (e.g. raw materials and components) and services firms acquire to external suppliers.

In (15.10) E' denotes the minimum amount of energy that a firm requires to produce output y, while holding constant its current level of non-energy inputs (K, L and M). Therefore, the first component in brackets in (15.11), the technical efficiency effect ($TECH_j$), is a ratio of two measures of technical efficiency that captures the rate at which a firm reduces (or increases) the waste in the use of energy in its existing production process. Note that energy savings can only arise from an improvement in the management of the use of energy, but not from any substitution between inputs because non-energy inputs are not allowed to vary.

By contrast, E^* denotes the minimum amount of energy that can be achieved among all technically feasible input combinations that permit obtaining a given level of output. In other words, E^* is the quantity of energy that results from the least energy intensive feasible input bundle to produce output level y. Consequently, in calculating E^* any input substitution possibilities are allowed, and firms are allowed to fully adjust K, L and M in any direction, i.e. the observed quantities of K, L and M can either decrease or increase. Consequently, the element $IMCH_j$ in (15.11) captures the contribution of the change in the input mix to the observed energy intensity change between periods t and $t + 1$. Specifically, a value of $IMCH_j$ lower (greater) than unity indicates that the input mix efficiency of firm

j has increased (decreased) between year 0 and 1, and thereby has contributed to reduce (increase) the observed energy intensity rate of firm j.

The distinction between E' and E^* is particularly relevant for the analysis of energy intensity because alternative combinations of non-energy inputs may result in substantial differences in energy consumption. On the one hand, the firm's energy consumption will be influenced by the capital-labor ratio K/L. In principle, one would expect that an increase of capital intensity would lead to higher energy consumption rates. For instance, the intensive use of equipment and the automation of production processes will typically require a higher amount of energy to produce certain goods than labor-intensive processes. In such case, it is said that capital and energy are complements. We note however that the effect of an increase of the K/L ratio can go either way, depending on the relationship between capital and energy (see e.g. Metcalf 2008; Song and Zheng 2012). Thus, if the increase of K is based on the replacement of capital stock by more energy efficient equipment then we should expect a decrease in the energy consumption. In this case energy and capital are substitutes.

On the other hand, for a given output level, the volume of the intermediate inputs M reflects the extent to which the activities are outsourced or conducted within the boundaries of the firm. In other words, it reflects the firm's buy or make decision, and thereby its degree of vertical integration. Let us consider two firms producing the same level of output but showing different levels of M. The firm with a low M would have internalized most of the value-creating activities associated with its business, consistent with a high level of vertical integration. In the case of a fully (perfectly) vertically integrated firm, $M = 0$. By contrast, a firm that procures most components and services externally would show a higher value of M and thus a lower degree of vertical integration. Broadly, we would expect that the lower the scope of vertical integration, the lower the energy intensity. In other words, a higher use of M is associated with a lower usage of both K and L, and thereby less energy consumption. Therefore, the ratio E'/E^* captures the energy savings that could be attained from changing the firm's degree of vertical integration and adjusting the capital-labor ratio.

Figure 15.2 illustrates the distinction between E' and E^*. The figure depicts a (piecewise linear) isoquant that shows the technically efficient combinations of two inputs, energy and other non-energy input, to produce a given level of output y. Point A therefore represents a firm that is technically inefficient. Thus, E' is the quantity of energy that results from the largest feasible contraction of the observed energy quantity, given the output level y, and the quantity consumed of other non-energy inputs x_a. To reach the isoquant at point B, firm A should reduce its excessive consumption of energy by the proportion E'/E_A.

In Fig. 15.2, company C has the input combination that uses the lowest quantity of energy to produce output level y, tough requiring a higher quantity of the other non-energy input x. If the company A seeks to reduce its energy consumption below E' it would require investing more on other inputs, up to x^*. Accordingly, the difference between E' and E^* accounts for the quantity of energy that could be saved if the company employs the lowest energy intensive input mix.

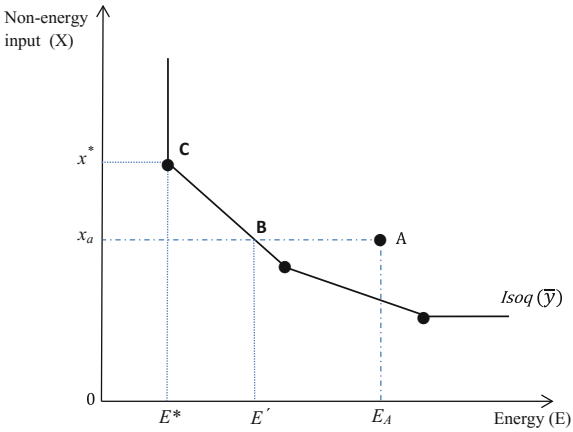


Fig. 15.2 The measurement of energy efficiency

15.3.2 The Sectoral Energy Intensity Change

Let us first consider the energy intensity in a particular industry. The total sectoral output is generated in a number of J economic units. In our empirical application regions are the economic units under consideration. Thus, the output and the energy consumption in industry i are respectively the sum of the output and the sum of the energy consumed in the J different regions that make up the industry i. Thus, the sectoral intensity can be expressed as a summation of the regional data:

$$\mathrm{I_{i,t}} = \frac{E_{it}}{Y_{it}} = \sum_{\mathrm{j}} \frac{\mathrm{E_{ij,t}}}{\mathrm{Y_{ij,t}}} \frac{\mathrm{Y_{ij,t}}}{\mathrm{Y_{it}}} = \sum_{\mathrm{j}} \mathrm{I_{ijt}} \mathrm{R_{ij,t}} \tag{15.12}$$

where E_{it} denotes the total energy consumption in industry i in period t, Y_{it} is the total output of industry i in period t; $E_{ij,t}$ is the quantity of energy consumed in region j in the within the industry i; $Y_{ij,t}$ is the output of industry i produced in region j in period t, and R_{ij} denotes the share of region j in the total output of sector i.

Let us assume that the energy intensity of industry i varies from $I_{i,0}$ in period 0 to $I_{i,1}$ in period 1. We apply the logarithmic mean Divisia index presented in Sect. 15.2 to multiplicatively decompose the sectoral energy intensity change as:

$$d\mathrm{I_{i,t}} = \frac{\mathrm{I_{i,1}}}{\mathrm{I_{i,0}}} = \left[\exp\left(\sum_j w_j \ln\left(\frac{I_{j,1}}{I_{j,0}}\right)\right)\right] \cdot \left[\exp\left(\sum_j w_j \ln\left(\frac{R_{j,1}}{R_{j,0}}\right)\right)\right] \tag{15.13}$$

where the summation is taken over the J regions, and

$$w_j = \frac{L\left(e_j^1, e_j^0\right)}{\sum_j L\left(e_j^1, e_j^0\right)} \tag{15.14}$$

In (15.14) $e_j = E_{i,j}/E_i$ is the share of region j on the total energy consumed in industry i, and L is the logarithmic mean function.

In expression (15.13) the change in sectoral energy intensity is expressed as the product of two elements. The first bracketed component in (15.13) captures the impact of the variation of regional energy intensity rates, while the second bracketed term captures the effect of the changes in the composition of the industry output. By introducing our decomposition of the regional energy intensity change as stated in (15.11) into the first bracketed term in (15.13) we obtain the decomposition of the energy intensity change in industry i as the product of five components:

$$\begin{aligned} dI_{i,t} = \frac{I_{i,1}}{I_{i,0}} = & \left[\exp\left(\sum_j w_j \ln(TECH_j)\right)\right] \cdot \left[\exp\left(\sum_j w_j \ln(IMCH_j)\right)\right] \\ & \cdot \left[\exp\left(\sum_j w_j \ln(TCH_j)\right)\right] \cdot \left[\exp\left(\sum_j w_j \ln(SCH_j)\right)\right] \\ & \cdot \left[\exp\left(\sum_j w_j \ln\left(\frac{R_{j,1}}{R_{j,0}}\right)\right)\right] \\ = & TECH_i \cdot IMCH_i \cdot TCH_i \cdot SCH_i \cdot REG_i \end{aligned} \tag{15.15}$$

The last component in (15.15) is the regional effect (REG_i) and measures the impact of the changes in the distribution of the sectoral output among regions. A value of REG_i lower (greater) than one indicates that production in industry i has moved from high (less) energy intensive regions to less (higher) energy intensive regions. The volume of production may increase in one region and decrease in others, and thereby increasing the share of the former on the total industry output. For instance, firms in one region may become more competitive and increase their production to the detriment of less competitive firms operating in other regions. Moreover, there are many reasons that may lead companies to move its activity from one region to another. For example, a specific region may offer economic advantages and more attractive conditions for business (e.g. lower labor costs, lower taxes, better infrastructures and services, higher availability of suppliers, etc.). However, while there are factors that give a firm some competitive advantage of operating in certain region, taking advantage of such factors may require, at the same time, to incur in higher energy needs (e.g. due to the new firm's location, weather, process organization).

15.3.3 The Aggregate Energy Intensity Change

The manufacturing industry is an aggregate comprising a wide range of economic activities. Thus, the overall energy consumption in manufacturing is defined as the

sum of the energy consumed in its $i = 1, \ldots, M$ different sectors. Hence, the aggregate energy intensity rate in period t can be expressed as

$$\mathrm{I}_t = \frac{E_t}{Y_t} = \sum_i \frac{\mathrm{E_{i,t}}}{\mathrm{Y_{i,t}}} \frac{\mathrm{Y_{i,t}}}{\mathrm{Y}_t} = \sum_i \mathrm{I_{i,t}\, S_{i,t}} \tag{15.16}$$

where E_t is the total energy consumption; $E_{i,t}$ is the energy consumption in industrial sector i; Y_t is the aggregate output; $Y_{i,t}$ is the output of industry i; $I_{i,t}$ is the energy intensity of sector i and $S_{i,t} = Y_{i,t}/Y_t$ is the production share of sector i.

By applying again the multiplicative LMDI decomposition the change in the aggregate energy intensity can then be expressed as

$$d\mathrm{I} = \frac{I_1}{I_0} = \left[\exp\left(\sum_i w_i \ln\left(\frac{I_{i,1}}{I_{i,0}}\right)\right)\right] \cdot \left[\exp\left(\sum_i w_i \ln\left(\frac{S_{i,1}}{S_{i,0}}\right)\right)\right] \tag{15.17}$$

where

$$w_i = \frac{L\left(e_i^1, e_i^0\right)}{\sum_i L\left(e_i^1, e_i^0\right)} \tag{15.18}$$

In (15.18) $e_i = E_{i,t}/E_t$ is the share of sector i in the aggregate energy consumption, and L is the logarithmic mean function defined above.

By introducing the decomposition of the sectoral energy intensity change as stated in expression (15.15) into (15.17), and denoting the variation in the output share of sector i as $SHARE_i = \frac{S_{i,1}}{S_{i,0}}$, we can express the full decomposition of the aggregate energy intensity change as

$$\begin{aligned} d\mathrm{I} = \frac{\mathrm{I_1}}{\mathrm{I_0}} = & \left[\exp\left(\sum_i w_i \ln(TECH_i)\right)\right] \cdot \left[\exp\left(\sum_j w_j \ln\left(IMCH_j\right)\right)\right] \\ & \cdot \left[\exp\left(\sum_i w_i ln(TCH_i)\right)\right] \cdot \left[\exp\left(\sum_i w_i \ln(SCH_i)\right)\right] \\ & \cdot \left[\exp\left(\sum_i w_i \ln(REG_i)\right)\right] \cdot \left[\exp\left(\sum_i w_i \ln(SHARE_i)\right)\right] \\ = & \, TECH \cdot IMCH \cdot TCH \cdot SCH \cdot REG \cdot STR \end{aligned} \tag{15.19}$$

The last element in expression (15.19) is the structural effect (*STR*), and captures the impact of the variation in the production structure on the aggregate energy intensity change. In other words, the structural change is associated with the varying growth rates among the constituent sectors of the aggregate industry, which lead to a change in its product mix. A value of *STR* lower than one indicates that production has shifted away from energy intensive industries.

15.4 Implementing the Decomposition of the Energy Intensity Rate Change

In the decomposition formulated in (15.11), E_t and y_t are respectively the energy and output quantities observed in period $t = 0,1$. However, the efficient energy quantities $E'_t, E^*_t\ E^*_t(y_{t+1}), E^*_{t+1}(y_t)$ are not directly observed and must be estimated from the observed data and technologies. Technologies are also unobserved, so they must be estimated too. A production technology transforms inputs $x = (x_1, \ldots, x_n)$ into outputs $y = (y_1, \ldots, y_m)$. The set of all input-output vectors that are feasible is called the production set (T), which is defined as

$$T = \left\{(x,y) \in \mathbb{R}^{n+m}_{+} : x \text{ can produce } y\right\} \tag{15.20}$$

We consider a piecewise linear sequential technology defined by the production set T^t as

$$T^t = \left\{(x,y) : y \leq \sum_{s=1}^{t}\sum_{j=1}^{J} z_j^s y_j^s, x \geq \sum_{s=1}^{t}\sum_{j=1}^{J} z_j^s y_j^s, z_j^s \geq 0, \sum_{s=1}^{t}\sum_{j=1}^{J} z_j^s = 1\right\} \tag{15.21}$$

The technology defined in (15.21) is constructed in a sequential way, by accumulating information from previous years for each of the J firms (Tulkens and Vanden Eeckaut 1995). Specifically, in the construction of period t technology we include data for all producers in t and all preceding years (from $s = 1$ to t). This approach implies the assumption that the way in which production has been performed in the past is always feasible for the company in subsequent years; in other words, technological regress is not possible. The convexity constraint $\{z_i^s 0, \sum_{z_i^s} = 1\}$ in expression (15.21) allows defining a production technology with variable returns to scale (Banker et al. 1984).

In our empirical application we assume that each firm produces only one output (m = 1) and employs four inputs (n = 4), being the input vector x = (x_K, x_L, x_E, x_M). The input efficiency measure necessary to calculate the energy quantity $E'_t(y_t)$ is calculated as the solution to the following linear programming problem:

$$
\begin{aligned}
&\min \theta \\
&\text{s.t.} \\
&y^t \le \sum_{s=1}^{t}\sum_{j=1}^{J} z_j^s y_j^s \\
&\sum_{s=1}^{t}\sum_{j=1}^{J} z_j^s x_{Ej}^s \le \theta x_E^t \\
&\sum_{s=1}^{t}\sum_{j=1}^{J} z_j^s x_{nj}^s = x_n^t \quad \text{n} = K, L, M \\
&\sum_{s=1}^{t}\sum_{j=1}^{J} z_j^s = 1 \\
&z_j^s \ge 0 \, \text{j} = 1, \ldots, J
\end{aligned}
\tag{15.22}
$$

The solution of (15.22) is an input subvector measure of technical efficiency as defined in Färe et al. (1994), where only the energy input x_E is scaled down and the other inputs are held constant at their observed levels. Thus, the efficient quantity of energy E_t' is given by

$$
E_t' = \theta \cdot E_t \tag{15.23}
$$

Similarly, $E_{t+1}'(y_{t+1})$ can be estimated as the solution to a linear programming problem identical to (15.22), just replacing period t data and technology with the corresponding $t + 1$ data and technology, i.e.

$$
\begin{aligned}
&\min \theta \\
&\text{s.t.} \\
&y^{t+1} \le \sum_{s=1}^{t+1}\sum_{j=1}^{J} z_j^s y_j^s \\
&\sum_{s=1}^{t+1}\sum_{j=1}^{J} z_j^s x_{Ej}^s \le \theta x_E^{t+1} \\
&\sum_{s=1}^{t+1}\sum_{j=1}^{J} z_j^s x_{nj}^s = x_n^{t+1} \quad \text{n} = K, L, M \\
&\sum_{s=1}^{t+1}\sum_{j=1}^{J} z_j^s = 1 \\
&z_j^s \ge 0 \quad \text{j} = 1, \ldots, J
\end{aligned}
\tag{15.22'}
$$

Finally, the maximum feasible shrinkage of the observed energy consumption (x_E) to produce y^t with period t technology is calculated as the solution to the following linear programming problem:

$$
\begin{aligned}
&\min \lambda \\
&\text{s.t.} \\
&y^t \le \sum_{s=1}^{t}\sum_{j=1}^{J} z_j^s y_j^s \\
&\sum_{s=1}^{t}\sum_{j=1}^{J} z_j^s x_{Ej}^s \le \lambda x_E^t \\
&\sum_{s=1}^{t}\sum_{j=1}^{J} z_j^s x_{nj}^s \ge 0 \quad \text{n} = K, L, M \\
&\sum_{s=1}^{t}\sum_{j=1}^{J} z_j^s = 1 \\
&z_j^s \ge 0 \quad \text{j} = 1, \ldots, J
\end{aligned}
\tag{15.24}
$$

Note that in (15.24) only the energy input (x_E) is scaled down, but unlike in (15.22), the use of capital (x_K), labor (x_L) and the intermediate inputs (x_M) is not constrained and can take any positive value. Consequently, the quantities of x_K, x_L, and x_M can decrease, increase or remain invariant with respect their observed level. Hence, the energy quantity E_t^* is the result of applying the largest feasible contraction of the energy input obtained from (15.24) to the observed energy quantity, i.e.

$$E_t^* = \lambda \cdot E_t \tag{15.25}$$

Similarly, $E_{t+1}^*(y_{t+1})$ can be estimated as the solution to a linear programming problem identical to (15.24), by just replacing period t data and technology with the corresponding $t + 1$ data and technology, i.e.

$$
\begin{aligned}
&\min \lambda \\
&\text{s.t.} \\
&y^{t+1} \le \sum_{s=1}^{t+1}\sum_{j=1}^{J} z_j^s y_j^s \\
&\sum_{s=1}^{t+1}\sum_{j=1}^{J} z_j^s x_{Ej}^s \le \lambda x_E^{t+1} \\
&\sum_{s=1}^{t+1}\sum_{j=1}^{J} z_j^s x_{nj}^s \ge 0 \quad \text{n} = K, L, M \\
&\sum_{s=1}^{t+1}\sum_{j=1}^{J} z_j^s = 1 \\
&z_j^s \ge 0 \quad \text{j} = 1, \ldots, J
\end{aligned}
\tag{15.24'}
$$

Finally, the energy efficiency measure needed to determine $E^{*}_{t+1}(y_t)$, i.e. the minimum feasible energy quantity to produce y^t with period $t + 1$ technology, is determined as the solution to a problem identical to (15.24), but now using as reference the period $t + 1$ technology:

$$
\begin{aligned}
&\min \lambda \\
&\text{s.t.} \\
&y^t \leq \sum_{s=1}^{t+1} \sum_{j=1}^{J} z_j^s y_j^s \\
&\sum_{s=1}^{t+1} \sum_{j=1}^{J} z_j^s x_{Ej}^s \leq \lambda x_E^t \\
&\sum_{s=1}^{t+1} \sum_{j=1}^{J} z_j^s x_{nj}^s \geq 0 \quad \text{n} = K, L, M \\
&\sum_{s=1}^{t+1} \sum_{j=1}^{J} z_j^s = 1 \\
&z_j^s \geq 0 \quad \text{j} = 1, \ldots, J
\end{aligned}
\tag{15.24''}
$$

15.5 Data and Variables

We apply the decomposition model shown above to quantify the impact of the different factors that explain the evolution of the energy intensity observed in the Spanish manufacturing between 1999 and 2007. The manufacturing output is disaggregated into nine sectors, which are those defined in the Energy Balances of the International Energy Agency for most countries. Table 15.1 shows the sectoral breakdown with its corresponding National Classification of Economic Activity (NACE) codes. To be precise, the nine sectors considered in Table 15.1 roughly accounts for 95% of the manufacturing output in Spain. We have left out the sectors that are traditionally assorted under the heading of "Other manufacturing industries" (NACE codes 22, 31 and 32), due to the lack of reliable information on energy consumption in those industries.

Table 15.1 Classification of manufacturing industries

	NACE codes	Energy intensity (E/Y)
Food, beverages and tobacco	10,11,12	0.045
Textile and leather	13,14,15	0.067
Wood and wood products	16	0.139
Paper, pulp and printing	17, 18	0.114
Chemical and pharmaceutical products	20, 21	0.123
Non-metallic mineral products	23	0.323
Basic metals	24	0.284
Machinery and equipment	25,26,27,28	0.019
Transport equipment	29, 30	0.016

The output of each industry is produced in seventeen regions. Nevertheless, we note that some less-industrialized regions show no production in certain industries. The sector with the smallest number of observations is the Transport equipment industry, whose total output is generated in thirteen regions.

We estimate separate production technologies for each industry from the observed input-output data to compute the energy efficiency measures for each region. As stated above, the frontier of each year is estimated sequentially from current and all previous (but not subsequent) data. In order to ensure a sufficient number of observations to estimate each of the annual production frontiers for the period 1999–2007, in the construction of the frontier corresponding to the first year (1999) we have accumulated information from the year 1994 to the year 1999.

In most analysis of energy intensity at macroeconomic level the production output is typically expressed in value added at basic prices. At macro level the costs of intermediate inputs cancel out against the gross income of delivering these inputs in the derivation of GDP. However, at industry level the intermediate deliveries do not cancel out. Thus, at industry level it is more appropriate the use of the gross production rather than value added as the output variable (EC 2014). Further, as Hulten (2010) argues, the use of value added as industrial output variable "implies (improbably) that efficiency-enhancing improvements in technology exclude material and energy" (Hulten 2010, p. 1004). Consequently, we use the gross production as the output variable in each industry.

In each region and industry the output is obtained from the utilization of four inputs: Capital, Labor, Energy and Materials. Data on Labor and Materials are readily available from the Industrial Companies Survey, conducted by the National Statistics Institute. Thus, labor quantity is measured by the number of worked hours, while Materials include the purchases of intermediate inputs and services consumed in the production process (excluding energy consumption) measured in constant 1995 €.

To obtain the quantities of capital and energy used in each region and sector we need to operate a little further. The energy consumption (in thousands of tons of oil equivalent, ktoe) in each industry is available only at national level, being provided

by the Spanish Institute for Diversification and Energy Savings (IDAE). By contrast, the energy expenditure is drawn from the Industrial Companies Survey both at sectoral and regional level. With this data we calculate the toes of energy consumed in each industry and region by proceeding in three steps. First, we draw from the Industrial Companies Survey the aggregate (national) expenditure on electricity, gas and other energies in each industry, and the quantities of final energy in physical units consumption in electricity, gas and other energies (mostly oil products and to a lesser extent, coal) from the IDAE. Secondly, we divide the total expense in each energy type and sector (electricity, gas and others) by its respective quantity consumed in physical units (ktoe). And so we get the national prices for each fuel and sector. We assume that within the same sector there are no differences in the price of electricity, gas and other energies across regions. Then, we divide the energy expenditure by its corresponding price to get the quantity of energy consumed in each sector and region.

For calculating the capital measure we also proceed in three steps. First, we extract from the Industrial Companies Survey the value of the depreciation expense for each region and industry. Second, we calculate the average depreciation rate applied in each industry from the data contained in the KLEMS database. Finally, we divide the depreciation expenses in each region by the sectoral average depreciation rate to get the regional capital stock in each sector.

The sectoral price indices employed to deflate the gross output and materials were obtained from the EU KLEMS database, while the price indices to deflate the capital stock series were drawn from the database provided by Fundación BBVA.

Table 15.1 shows the mean values of the energy intensity by sector, revealing substantial differences in the energy intensity across industries. Thus, the most energy intensive industries are the Non-metallic mineral products and the Basic Metals, which use much more energy than the other industries to produce one unit of output. On the contrary, the Transport equipment and Machinery & equipment sectors are the less energy intensive consumers.

15.6 Results

Table 15.2 shows the results of our decomposition of the energy intensity change for the entire manufacturing industry between 1999 and 2007. First column in Table 15.2 reports the observed annual change in energy intensity, while columns (2)–(7) show the yearly changes for the six components identified in Eq. (15.19). Last row in each column shows the change rate of every factor in cumulative terms from 1999 to 2007. Figure 15.3 displays the evolution of the cumulative change in the energy intensity and its components over the period under consideration.

The number at the bottom of the first column in Table 15.2 is 0.879, indicating that the energy used per unit of output obtained in the aggregate manufacturing industry decreased by 12.1% between 1999 and 2007. A look at Fig. 15.3 reveals that the energy intensity presents a cumulative growth until 2005, showing an

Table 15.2 The energy intensity change and its components (total manufacturing sector)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	*d*I	TECH	IMCH	TCH	SCH	REG	STR
1999/00	1.061	0.985	1.035	0.930	1.096	1.000	1.022
2000/01	1.026	1.041	0.954	0.975	1.032	1.000	1.027
2001/02	0.986	0.998	1.001	0.949	1.015	1.010	1.014
2002/03	1.073	0.976	1.120	0.993	1.003	1.003	0.983
2003/04	0.983	0.979	0.949	0.979	1.046	1.002	1.031
2004/05	0.967	0.997	0.968	0.955	1.032	0.999	1.016
2005/06	0.816	0.971	0.883	0.902	1.024	1.003	1.028
2006/07	0.985	1.046	0.984	0.959	0.999	1.001	0.997
1999/07	0.879	0.990	0.884	0.691	1.271	1.018	1.123

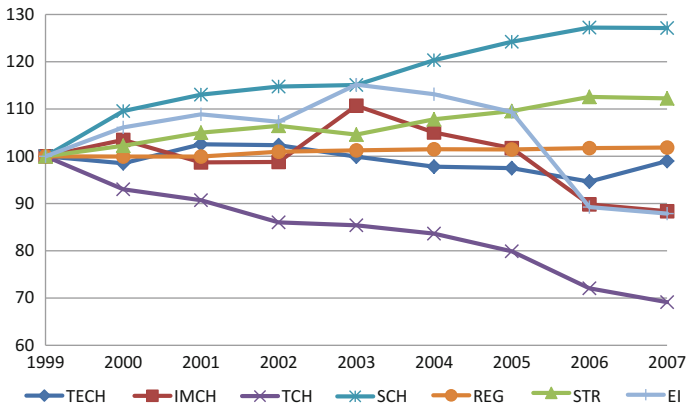


Fig. 15.3 The energy intensity change of manufacturing and its components (1999–2007)

abrupt reduction in 2006. A similar sharp decline in 2006 is reported in the official statistics (MICYT 2006) relative to the overall Spain's energy intensity (measured in toe per GDP).

As shown in column (4), such decline is primarily due to Technical Change, which would have reduced the energy intensity rate by 30.9%. Figure 15.3 confirms that a steady and significant technological progress occurred throughout the period. The second most important factor in reducing energy intensity has been the change in the input mix (IMCH). Thus, column (3) indicates that this effect would have led to a decrease of 11.6% in the manufacturing energy intensity. The results of the effect of the Technical Efficiency Change (EMCH) in column (2) suggests that only a slight improvement in the use of energy occurred within firms, which would have caused a cumulative positive effect of 1%.

However, the energy reducing effect of the aforementioned factors is partially offset by the evolution of the other three components. Above all others, the scale

effect (SCH) reported in column (5) is notably higher than one, indicating that the increase in the scale of operations registered during the period in the manufacturing sector was accompanied by a much more intensive use of energy. Specifically, the energy intensity of manufacturing would have increased by 27.1% throughout the period due to the scale effect. Figure 15.3 displays the consistent and increasing trend of the scale effect since the beginning of the period under consideration.

Secondly, column (7) shows the impact of the change registered in the sectoral composition of the Spanish manufacturing industry throughout the analyzed period. Specifically, the Structural Effect (STR) would have contributed to rise the aggregate energy intensity of manufacturing by 12.3%, motivated by the increase of the share that the relatively higher energy-intensive activities have in the aggregate industrial output.

Thirdly, the Regional Effect (REG) in column (6) indicates that the variation in the distribution of the output across regions that occurred over the considered period would have increased the aggregate energy intensity by 1.8%.

Thus far we have discussed the general results for the aggregate manufacturing. Let us analyze now the results of individual industries. First of all, before discussing the results relative to the decomposition of the sectoral energy intensity change, we focus on the analysis of the level energy efficiency at sectoral level. Specifically, Table 15.3 shows the estimates of the two energy efficiency measures defined by the two linear programming problems (15.22) and (15.23), as well as the differences in the composition of the input vectors corresponding to both energy efficiency references. All figures in Table 15.3 are the mean values registered over the period 1999–2007.

The first column in Table 15.3 shows the value of the technical energy efficiency achieved in each industry over the period under consideration. That is, it shows the average value of θ that results from solving the linear programming problem written in (15.22) for every year. For instance, we note that the average technical energy efficiency in the Food industry is 0.848, indicating that the Food sector could reduce its consumption of energy by 15.2% without reducing output and holding fixed the observed levels of the other inputs.

Table 15.3 Energy efficiency

	$\theta = E'/E$	$\lambda = E^*/E'$	M^*/M	K^*/M	L^*/L
Food	0.848	0.754	1.008	0.671	0.867
Textile	0.961	0.800	1.017	0.764	0.963
Paper	0.777	0.562	0.977	0.751	1.154
Chemical	0.814	0.630	0.998	0.751	1.057
Non-metallic mineral products	0.934	0.737	1.057	0.782	0.933
Transport equipment	0.808	0.547	1.024	0.822	0.857
Wood	0.872	0.640	1.012	0.687	1.130
Basic metals	0.931	0.580	1.043	0.901	1.154
Machinery and equipment	0.885	0.663	1.020	0.696	0.834

Mean values (1999–2007)

The second column in Table 15.4 lists the value of the input mix energy efficiency, i.e. the average value of λ that results from solving the linear programming problem written in (15.23) for every year. In the case of the Food industry the value of λ is 0.754. That is, the Food sector could further reduce the consumption of energy by 24.6%, if the optimal least energy demanding input mix is adopted, without reducing output.

The Textile industry appears to be the industry that achieves the highest energy efficiency, both regarding the technical efficiency (0.961) and the input mix efficiency (0.800). Conversely, our results suggest that in the Paper industry there is a large room for improvement in both efficiency dimensions.

Last three columns in Table 15.3 compare the energy-efficient input quantities associated to $E^*(K^*, L^*, M^*)$ with the observed input quantities (K, L, M), which we recall are held fixed in computing the technically efficient energy quantity E'. That is, the ratios between the two show the changes that every sector should make in their input bundle to achieve the smallest energy consumption. In the case of the Food industry, the optimal mix would require a quantity of purchased intermediate inputs 0.8% greater than the quantity actually consumed, while using 32.9% less capital, and 13.3% less labor than their respective observed quantities.

Most industries reveal a fairly similar pattern: to achieve the smallest energy consumption it is required to increase the purchase of intermediate inputs and decrease the use of capital in a larger proportion than labor, resulting in a lower capital intensity of production. There is however some interesting exceptions. In some industries the increase of M is accompanied by an increase of labor (e.g. Basic Metals, Wood). In the Paper industry, and to a lesser extent in the Chemical sector, the energy efficient vector is associated with fewer purchases of materials and services. This suggests that in these manufactures, the processing of purchased raw materials and components needs more energy than the processing of internally made feedstock.

Table 15.4 and Fig. 15.4 present the energy intensity change and its determinants for the nine manufacturing sectors considered in this study. In Table 15.4, the

Table 15.4 The energy intensity change and its components across industries 1999–2007

	(1)	(2)	(3)	(4)	(5)	(6)
	TECH	IMCH	TCH	SCH	REG	dI
Food	0.910	0.919	0.877	1.067	1.014	0.793
Textile	1.032	1.023	0.728	0.875	0.907	0.610
Paper	1.142	0.829	0.623	1.425	1.025	0.861
Chemical	1.320	0.807	0.817	1.213	1.040	1.097
Non-metallic mineral products	0.982	0.880	0.653	1.333	1.019	0.767
Transport equipment	0.824	0.953	0.884	1.225	0.984	0.837
Wood	0.725	1.327	0.545	1.512	1.043	0.827
Basic metals	0.869	0.846	0.553	1.406	1.013	0.579
Machinery and equipment	1.009	0.919	0.942	1.066	1.058	0.984

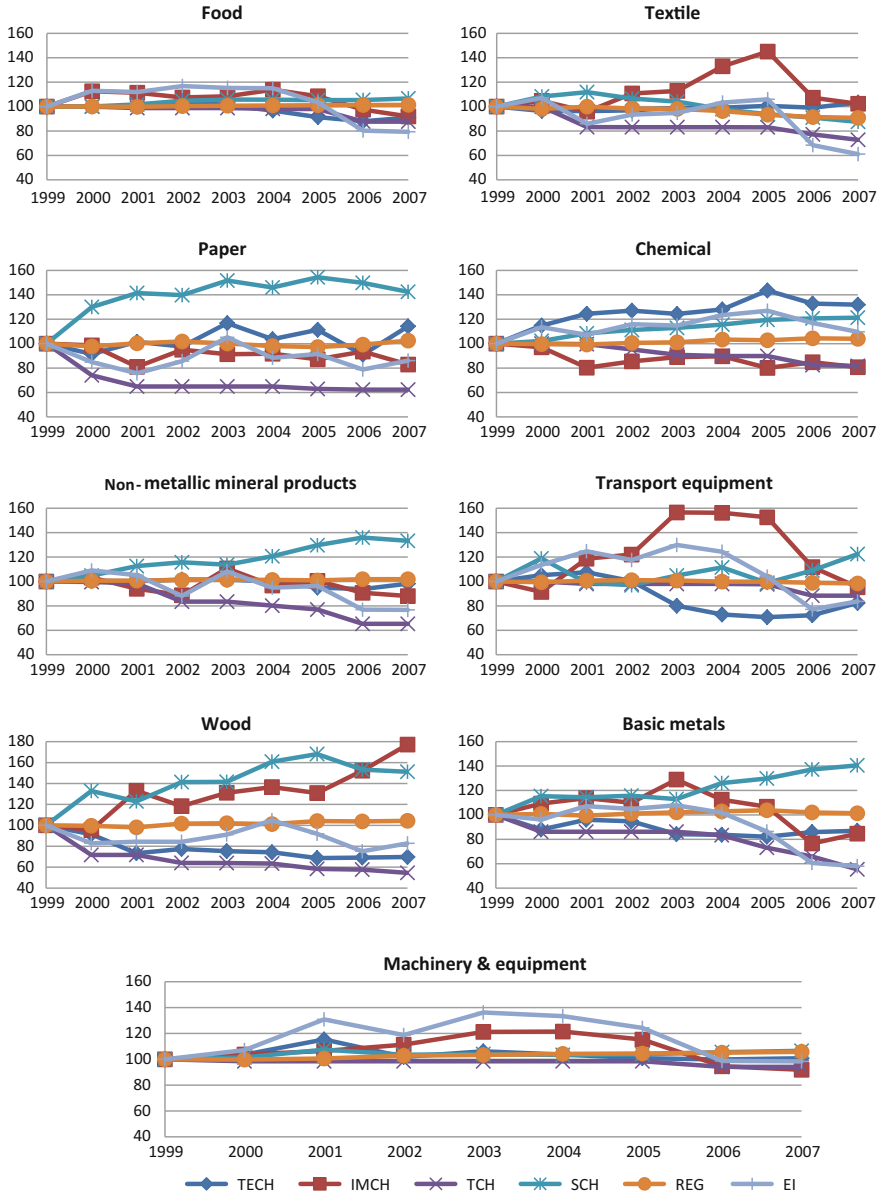


Fig. 15.4 The cumulative energy intensity change and its components

horizontal product of columns (1)–(5) equals the energy intensity change observed in 1999–2007 for each sector, which is reported in column (6).

Last column of Table 15.4 reveals that eight out of nine sectors improved its energy intensity over the period under consideration. Only the Chemical sector

registered an increase of 9.7% in its energy intensity rate. The Basic Metals industry shows the largest reduction (42.1%), followed by the Textile industry (39%). Table 15.4 nevertheless shows notable disparities across sectors with respect to the sign and magnitude of the explanatory factors.

As shown in column (3), Technical Change is the main source of energy intensity reduction in most sectors. Nevertheless, in two industries (Chemical and Machinery & Equipment) the effect of the input mix change is the most important driver, as can be seen in column (2). Column (1) reveals that the improvement of technical energy efficiency is the dominant energy-reducing factor in the Transport Equipment industry, while Wood, Basic Metals and Food sectors also show important savings derived from this efficiency increase.

As column (4) indicates, the scale change effect is the main negative force working against energy intensity. The sole exception is the Textile industry, where the change in the scale of operations had a positive impact of 12.3%. The period analyzed was a time of high sustained growth in Spain. The significantly negative scale effect detected in various industries indicates that such output increase required a proportionally higher consumption of energy, suggesting the presence of firm size inefficiencies within those industries with respect to the use of energy. Finally, with the exceptions of the Textile and Transport equipment sectors, the regional effect has contributed to deteriorate the energy intensity change. This suggests that in seven out of nine industries production has moved from less energy intensive regions to higher energy intensive regions.

15.7 Conclusions

This chapter has presented a way to analyze the change in energy intensity that combines frontier methods and the index decomposition approached usually employed in energy studies. The suggested decomposition has the advantage of providing a more detailed number of determinants as well as allowing an integrated analysis of the relationship between energy efficiency and energy intensity.

We have applied the proposed decomposition to the analysis of the evolution of energy intensity in Spanish manufacturing over the period 1999–2007 by using regional and industry level data. Broadly, our findings confirm that the technical progress, the change in the input mix and the improvement in the level of technical energy efficiency are the factors that have contributed to reduce the energy intensity in most manufacturing industries throughout the analyzed period. By contrast, the increase in the scale of operations and the change in the regional distribution of production have acted as energy intensity increasing forces within most industries.

In any case, the individual results at regional level should be interpreted with caution due to the level of aggregation employed in defining the industries. Undoubtedly, the accuracy of frontier estimation and efficiency measures would be higher if we could observe data at the four-digit NACE level and give a more homogeneous definition of each industry.

Finally, in Spain, as in many other countries, the enhancement of energy efficiency and the reduction of energy consumption represent major economic and environmental challenges, as reflected in the successively approved National Energy Efficiency Action Plans (MITYC 2007, 2011). In such a context, our analysis can be of help to the industrial policy assessment by identifying the driving forces that contribute to decline the energy intensity at industry level, and thereby guiding policy makers in the design of alternative measures and incentives to further reduce the energy consumption in different industries. Thus, in the light of our results, in some industries (e.g. Textile) bringing policy measures aimed at incentivizing a better energy management and the adoption of changes in their capital-labor ratio would be particularly suitable to reduce the consumption of energy. By contrast, in other industries (e.g. Non-metallic mineral products, Basic metals) the application of a different package of measures are expected to be more effective (e.g. giving stronger incentives to increase the production in smaller firms, to introduce changes in the degree of vertical integration, to stimulate the production in certain regions).

Acknowledgements The authors acknowledge financial support from the Spanish Ministry of Economy and Competitiveness (research project ECO2013-46954-C3-1-R).

References

Ang BW, Liu FL (2001) A new energy decomposition method: perfect in decomposition and consistent in aggregation. Energy 26:537–548

Ang BW, Zhang FQ (2000) A survey of index decomposition analysis in energy and environmental studies. Energy 25:1149–1176

Ang BW (1995) Decomposition methodology in industrial energy demand analysis. Energy 20 (11):1081–1095

Ang BW (2004a) Decomposition analysis applied to energy. Encycl Energy 1:761–769

Ang BW (2004b) Decomposition analysis for policymaking in energy: which is the preferred method. Energy Policy 32:1131–1139

Ang BW (2015) LMDI decomposition approach: a guide for implementation. Energy Policy 86:233–238

Ang BW, Choi KH (1997) Decomposition of aggregate energy and gas emission intensities for industry: a refined Divisia index method. Energy Journal 18(3):59–73

Ansuategi A, Arto I (2004) La evolución de la intensidad energética de la industria vasca entre 1982 y 2001: un análisis de descomposición. Economía Agraria y Recursos Naturales 4 (7):63–91

Balk BM (2008) Price and quantity index numbers: models for measuring aggregate change and difference. Cambridge University Press, New York

Balk BM (2010) An assumption-free framework for measuring productivity change. Rev Income Wealth 56(Special issue 1): S224–S256

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30:1078–1092

Boyd GA, Hanson DA, Sterner T (1988) Decomposition of changes in energy intensity: A comparison of the Divisia index and other methods. Energy Econ 10(4):309–312

Cansino JM, Sánchez-Braza A, Rodríguez-Arévalo M (2015) Driving forces of Spain's CO_2 emissions: a LMDI decomposition approach. Renew Sustain Energy Rev 48:749–759
Colinet MJ, Román R (2015) LMDI decomposition analysis of energy consumption in Andalusia (Spain) during 2003-2012: the energy efficiency policy implications. Energ Effi. doi:10.1007/s12053-015-9402-y
Deloitte (2010) Global manufacturing competitiveness index. Deloitte Touche Tohmatsu (DTT) Global Manufacturing Industry and the United States Council on Competitiveness
EC (2006) Directive 2006/32/EC of the European Parliament and of the Council on Energy End-Use Efficiency and Energy Services and Repealing Council Directive 93/76/EEC. European Parliament and Council, April 2006
EC (2010) Europe 2020: a European strategy for smart, sustainable and inclusive growth. European Commission, 3/3/2010
EC (2012) Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on Energy Efficiency, Amending Directives 2009/125/EC and 2010/30/EU and Repealing Directives 2004/8/EC and 2006/32/EC. Off J Eur Union L 315:1–56, 14 Oct 2012
EC (2014) Energy economic developments in Europe. European Economy 1. European Commission, Brussels, Jan 2014
Färe R, Grosskopf S, Lovell CAK (1994) Production frontiers. Cambridge University Press, Cambridge
Farrell MJ (1957) The measurement of productive efficiency. J R Statist Soc Ser A Gen 120 (3):253–282
Fernández P, Landajo M, Presno MJ (2013) The Divisia real energy intensity indices: evolution and attribution of percent changes in 20 European countries from 1995 to 2010. Energy 58:340–349
Fernández P, Landajo M, Presno MJ (2014) Multilevel LMDI decomposition of changes in aggregate energy consumption. A cross country analysis in the EU- 27. Energy Policy 68:576–584
Fernández-González P, Pérez-Suárez R (2003) Decomposing the variation of aggregate electricity industry in Spanish industry. Energy 28:171–184
Filippini M, Hunt LC (2011) Energy demand and energy efficiency in the OECD countries: a stochastic demand frontier approach. Energy J 32:59–80
Filippini M, Hunt LC (2012) US residential energy demand and energy efficiency: a stochastic demand frontier approach. Energy Econ 34:1484–1491
Filippini M, Hunt LC (2015) Measurement of energy efficiency based on economic foundations. Working Paper 15/216. CER-ETH-Center of Economic Research at ETH Zurich. Swiss Federal Institute of Technology, Zurich
Hoekstra R, van den Bergh JCJM (2003) Comparing structural and index decomposition analysis. Energy Econ 25(1):39–64
Hulten CR (2010) Growth Accounting. In: Hall BH, Rosenberg N (eds) Handbook of the economics of innovation, vol 2. Elsevier, Amsterdam, pp 987–1031
IEA (2013) World energy outlook 2013. International Energy Agency, Paris
Marrero GA, Ramos-Real FJ (2008) La intensidad energética en los sectores productivos en la UE-15 durante 1991 y 2005: ¿Es el caso español diferente? Estudios Económicos FEDEA 08-08
Mendiluce M, Pérez-Arriaga JI, Ocaña C (2010) Comparison of the evolution of energy intensity in Spain and in the EU15. Why is Spain different? Energy Policy 38(1):639–645
Mendiluce M (2007) Cómo afectan los cambios estructurales a la intensidad energética en España, Ekonomiaz, 65, 2º cuatrimestre: 362–385
Metcalf GE (2008) An empirical analysis of energy intensity and its determinants at the state level. Energy J 29(3):1–26
MITYC (2006) La Energía en España. Ministerio de Industria Turismo y Comercio, Madrid
MITYC (2007) Plan de Acción de Ahorro y Eficiencia Energética en España 2004-2012 (E4). Instituto para la Diversificación y Ahorro de la Energía. Madrid, Ministerio de Industria Turismo y Comercio
MITYC (2011) Plan de Ahorro y Eficiencia Energética 2011-2020. Instituto para la Diversificación y Ahorro de la Energía. Madrid, Ministerio de Industria Turismo y Comercio

Mulder P, De Groot HLF (2012) Structural change and convergence of energy intensity across OECD countries, 1970-2005. Energy Econ 34:1910–1921

Orea L, Llorca M, Filippini M (2015) A new approach to measuring the rebound effect associated to energy efficiency improvements: an application to the US residential energy demand. Energy Econ 49:599–609

Sato K (1976) The ideal log-change index number. Rev Econ Stat 58:223–228

Song F, Zheng X (2012) What drives the change in China's energy intensity: combining decomposition analysis and econometric analysis at the provincial level. Energy Policy 51:445–453

Su B, Ang BW (2012) Structural decomposition analysis applied to energy and emissions: some methodological developments. Energy Economics 34:177–188

Tulkens H, Vanden Eeckaut P (1995) Non-parametric efficiency, progress and regress measures for panel data: methodological aspects. Eur J Oper Res 80(3):474–499

Vartia YO (1976) Ideal log change index numbers. Scand J Stat 3:121–126

Voigt S, De Cian E, Schymura M, Verdolini E (2014) Energy intensity developments in 40 major economies: structural change or technology improvement? Energy Econ 41:47–62

Zhou P, Ang BW (2008) Linear programming models for measuring economy-wide energy efficiency performance. Energy Policy 36:2911–2916

Zhou P, Ang BW, Poh KL (2008) A survey of data envelopment analysis in energy and environmental studies. Eur J Oper Res 189:1–18

Chapter 16
The Aging U.S. Farmer: Should We Worry?

Harold O. Fried and Loren W. Tauer

Abstract The average age of the U.S. farmer continues to increase and exceeded 58 years in 2012. This aging of farmers is not unique to the U.S. If older farmers are less productive than younger farmers then agricultural output may diminish. Although Malmquist techniques are often used to measure productivity over time, we measure the productivity of farmers of various age cohorts with DEA techniques using the 50 state-level age cohort data from the 2012 U.S. Agricultural Census. We define a global technology comprising data from all age groups and states. Productivity of an age cohort in a state is then measured relative to data from all age groups rather than between adjacent ages. The efficiency component of a state age group is measured relative to the other state observations in that age group. Technology is measured as the Malmquist productivity value divided by efficiency. We find that the productivity of the age group of 35–44 years old is 3% more productive than the youngest farmers under the age of 25, but that the productivity of farmers over the age of 65 is 10% lower than the youngest farmers. The decrease in productivity of old farmers is due to technology because on average they remain efficient.

Keywords Age productivity · DEA · Farmer productivity · Global technology · Malmquist

An early draft of this paper was presented at the International Workshop on Efficiency and Productivity in June, 2015 in Alicante, Spain, sponsored by the University Research Institute "Center of Operations Research," University Miguel Hernandez of Elche. We thank the participants for a productive discussion, stimulated by the fine weather and hospitality. An anonymous referee provided thoughtful and insightful comments and suggestions that transformed the paper, including a new methodology.

H.O. Fried (✉)
Department of Economics, Union College, Schenectady, NY, USA
e-mail: friedh@union.edu

L.W. Tauer
Charles H. Dyson School, Cornell University, Ithaca, NY, USA
e-mail: lwt1@cornell.edu

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7_16

16.1 Introduction

This paper is about the productivity of older farmers. The concern is that older farmers may be less productive than younger farmers, and U.S. farmers on average are becoming older. If older farmers are less productive and if farmers continue to age, future farm productivity will decrease. This has important policy implications: programs could be put into place that mitigate productivity decreases with age, or to encourage the transition of farm businesses from older and less productive farmers to younger and more productive farmers. These issues have been discussed by U.S. Secretary of Agriculture, Tom Vilsack. They are also an issue in other countries where agriculture is a dominant industry, such as New Zealand (Fairweather and Mulet-Marquis, 2009).

Aging and productivity is an issue that applies across individuals and across occupations. As workers get older, they gain experience, which contributes to productivity, but beyond some point, the physical and mental deterioration that is inherent in the aging process manifests itself as a countervailing and eventually prevailing force to the positive contribution of knowledge gained on the job. The result is an inverted "U" shaped relationship between age and productivity as the experience effect dominates initially, only to be trumped by the aging effect eventually. The manifestation of any upside down "U" is expected to vary across individuals and occupations. Opportunities exist for economists and other social scientists to investigate this relationship empirically.

The literature on aging and productivity adopts various perspectives: the relationship between aging, productivity and wages; the management of older workers; the time path of mental and physiological aging and productivity; managing the aging process; aging and economic growth. We focus on aging and productivity.

Feyrer (2007) investigates the relationship between workforce demographics and economic growth using a large panel of OECD and developing countries. He concludes that countries with older demographics tend to grow faster than countries with younger demographics; more workers between the ages of 40–49 is associated with higher growth. Skirbekk (2003) surveys the literature on age and individual productivity. In general, job performance tends to peak around age 50; specifically, performance in jobs that involve problem solving, learning and speed, experience more rapid deterioration with age than jobs that involve experience and verbal abilities. Oster and Hamermesh (1998) examine economists' publishing in leading journals over their careers and find evidence that performance deteriorates with age: "creative economics at the highest levels is mainly for the young." Turning to two examples from sports, Fried and Tauer (2011, 2012) investigate aging and golf for men on the PGA tour and women on the LPGA tour. The performance metric is the ability to perform under pressure, which peaks at age 36 for the men and age 37 for the women.

The life cycle pattern of farmer productivity is a testable hypothesis. Tauer (1984, 1995) investigates this pattern using data from various U.S. Agriculture Census years. These analyses are across age groups at a point in time rather than

following farmers as they age. He finds evidence that farmer productivity with respect to age exhibits first an increase and then a decrease across farmer age cohorts at various Census time periods. Tauer and Lordkipanidze (2000) investigate the sources of productivity change, decomposing it into technology and efficiency changes. This matters since policies to boost efficiency are different from policies to encourage the adoption of new technology. Understanding the profile of efficiency change and technology change with respect to age enables policy makers to target policy interventions to farmers of different ages. Mishra and El-Osta (2008), for instance, find that farm succession decisions are significantly influenced by government farm policy. Like most industries, agriculture has changed, so we revisit this phenomenon using data from the most recent 2012 U.S. Census of Agriculture.

There are numerous studies that estimate and decompose productivity change in agriculture in countries around the world, often calculating Malmquist indices using a non-parametric approach. However, productivity differences across age cohorts has received almost no attention despite the fact that farmers are getting older all over the world. Gale (1994) does use Agricultural Census data from the years of 1978, 1982, and 1987 to study farm patterns over time and age, and although he does not estimate productivity by age, he does find that mean growth rates are greatest for younger farmers. Katchova and Ahearn (2015) focus on farm expansion rather than productivity and find that younger beginning farmers tend to expand over time in contrast to older beginning farmers. Our paper calculates a Malmquist index using a cross section methodology for 2012 and decomposes it into efficiency and technology, and further components of efficiency.

As U.S. farms become larger through consolidation, fewer farmers of all ages are needed. Because fewer additional farmers are needed after consolidation, it is logical that fewer younger individuals may enter farming, skewing the age distribution to older farmers and increasing the average age. The appeal of farming to the young may also be waning so that older farmers remain farming, given no apparent successors, although these operations will eventually be consolidated.

Principal operators are generally getting older each Census, exceeding 58 years in the 2012 Census.[1] See Fig. 16.1. The age reported is the age of the principal operator of sole-managed or multi-managed farms. The other managers may be younger, or even older than the reported principal operator. The census form is sent to each farm, and the farm determines the principal operator. With few exceptions these are primarily family farms.

Young farmers often do not survive as farmers. Gale (1994) studied the age of farmer exits and entries and concluded that over the period 1978–1997, the number of entries by young farmers declined steadily as did the exit rate of older operators, increasing the average farmer age in U.S. agriculture. This pattern persists in later censuses. Comparing principal operators by tenure on the farm in the 2007 and

[1]Over the previous 50 plus years the only time the average farmer age decreased was during the high commodity price years from 1973 to 1982 when high farm incomes encouraged young individuals to enter farming. Many of them may have been lost to farming during the farm financial crisis of the 1980s.

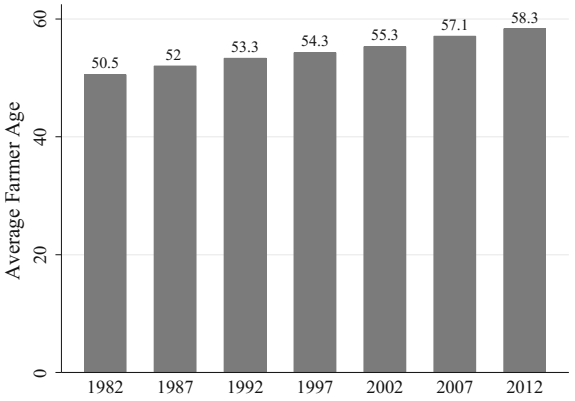


Fig. 16.1 Average age of the Principal Farm Operator, U.S. Agriculture Census Years 1982–2012

2012 Census, farmers with tenure of 9 years and less are leaving farming; farmers with tenure ten years and more are growing (Kurtzleben 2014). It is clear that the average farmer is older today and if this long term trend continues, the average farmer will be older in the future.

The remainder of this paper is organized as follows. The next section discusses the Malmquist methodology using Data Envelopment techniques to measure and decompose productivity into efficiency, and technology components. The standard Malmquist approach is typically applied to panel data. In this paper, we modify the standard approach to apply it to a cross section. Section 16.3 presents the results. The final section concludes.

16.2 The Method: Cross Section Malmquist with a Global Technology Set

16.2.1 Standard Malmquist

The Malmquist productivity index, typically based upon panel data of DMUs over time, is calculated at the level of the individual unit and can be decomposed into efficiency and technology effects and further into scale and pure (residual) components of efficiency (Fare, Grosskopf, and Margaritis 2008). Between two time periods, a firm undergoes productivity change as a result of a movement closer or further from the shifting production frontier (efficiency change), an improvement or deterioration as a result of adopting or failing to adopt new technology (technology change), and moving to a point of increasing or decreasing returns to scale (efficiency scale effect).

Productivity growth for an individual unit is the change in output over input between two time periods. In our context, this emanates from three sources: a

change in the position of the firm due to an expanding production possibilities set (technology), a change in the position of the firm due to a change in scale of production (scale efficiency), and a change in the position of the firm relative to the constant return production possibilities set (pure efficiency). Productivity change between two time periods is the sum of the three.

16.2.2 Malmquist with a Twist

This paper applies the Malmquist methodology to cross sectional data. The objective is to identify productivity differences among farmers of different age cohorts at a point in time, whereas the objective of traditional Malmquist is to identify productivity change between two adjacent time periods. Productivity change can still be decomposed into efficiency and technology components. The farmer age cohort dimension substitutes in the cross section context for the time dimension in the panel context. An advantage of cross section Malmquist is that it holds constant variables that change over time. This is particularly important for agriculture where changes in the weather over time can confound changes in productivity.

However, productivity can be measured relative to an adjacent age group, which is analogous to measuring productivity over time by measuring the change in productivity between adjacent time periods, or productivity of an age cohort can be measured relative to the technology of all age cohorts. This technique of specifying a global reference set for measuring productivity rather than an adjacent technology set was introduced by Pastor and Lovell (2005) and used by Camanho and Dyson (2006). Whereas the earlier measurement of farmer age productivity by Tauer and Lordkipanidze (2000) used adjacent age cohorts, in this paper we elect to use the Pastor and Lovell (2005) global specification of the technology set over all age cohorts. A recent application of Malmquist using a global technology set is Asmild (2015). A global specification is appropriate because we are interested in the productivity of a farmer age cohort relative to the most productive farmer regardless of the age of that productive farmer. This permits productivity measurement of any specific age group relative to the most productive group and allows identification of the most productive age group. Although we use this procedure to measure productivity across age groups, the method can also be used to measure productivity differences across groups of regions or groups of industries.

An output distance function can be defined for age group k as:

$$D_o^k\left(x^k, y^k\right) = \left(\max\left\{\theta: \left(x^{k'}, \theta y^{k'}\right) \in s^k\right\}\right)^{-1}. \tag{16.1}$$

This measures how much output y can be increased for decision maker k' given a quantity of input(s) x used by k', such that x and θy remain in the production set s^k.

An output rather than an input distance function is used because farmers are more likely increase their outputs given their use of inputs, rather than decrease inputs given their outputs. The function D_o^k measures the output technical efficiency of age cohort k, given the technology set used by the members of age cohort k. The technical efficiency difference between any two age cohorts is then

$$\mathrm{E}_{\mathrm{o}}^{\mathrm{k}+\mathrm{j}}\left(y^{k+j}, x^{k+j}, y^k, x^k\right) = \frac{D_o^{k+j}\left(x^{k+j}, y^{k+j}\right)}{D_o^k\left(x^k, y^k\right)}, \tag{16.2}$$

where k is age group k, referred to as the base age group, and j is age group j. If adjacent age groups are used than j = 1.

To construct the Malmquist index from a global reference technology, it is necessary to define distance functions for members of an age cohort in reference to the combined age cohort technology set as:

$$G_o^k\left(x^G, y^G\right) = \left(\max\left\{\theta: \left(x^{k'}, \theta y^{k'}\right) \in s^G\right\}\right)^{-1}, \tag{16.3}$$

where the G superscript refers to the global inputs x and outputs y from the combined age cohorts, which is the union of the various age cohorts technology sets. The output distance function specified by Eq. (16.3) measures the maximal proportional change in output required to make ($x^{k'}$, $y^{k'}$) feasible in relation to the global technology set s^G used by the combined age groups, and is the defined Malmquist index. The Malmquist index change between any two age groups is then:

$$M_o^{k+j}\left(x^G, y^G\right) = G_o^{k+j}\left(x^G, y^G\right) / G_o^k\left(x^G, y^G\right), \tag{16.4}$$

where k is the base age group and j is the age group being evaluated. Again, if j = 1 then the Malmquist index is being computed for adjacent age groups.

By definition the Malmquist index consists of the product of technology and efficiency. Efficiency of a DMU is measured by Eq. (16.1) and the difference in efficiency between adjacent age cohorts is measure by Eq. (16.2). Thus technology difference between any two age groups can be obtained by dividing Eq. (16.4) by Eq. (16.2): $T_o^{k+j}(.) = M_o^{k+j}(.)/E_o^{k+j}(.)$, where again if j = 1 then adjacent age groups are used in this evaluation.

Because Malmquist is not transitive across time periods (Førsund 2002) or in our case across age groups, we elect to measure the Malmquist, efficiency, and then technology of every age group relative to the age cohort of farmer under the age of 25. By using this constant base all indices by definition are transitive to that base. Thus the Malmquist, efficiency, and technology indices of all farmers in all age cohorts are measured relative to the farmers under 25 years of age. This also converts these three indices to the value of one for all farmers under the age of 25.

The defined distance functions can be calculated for each age group using linear programming techniques. The linear programming model to calculate the output distance function (16.1) for each of the k′ state age cohorts in age cohort group k is:

$$\left(D_o^k\left(x^{k'}, y^{k'}\right)\right) = \max \theta^{k'} \tag{16.5}$$

subject to:

$$\begin{aligned} &\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{z}^{\mathrm{k}} y_m^k \geq \theta^{k'} y_m^{k'} \quad \mathrm{m} = 1, \ldots, \mathrm{M} \\ &\sum_{k=1}^{K} z^k x_n^k \leq x_n^{k'} \quad \mathrm{n} = 1, \ldots, \mathrm{N} \end{aligned} \tag{16.5a}$$

$$z^k \geq 0 \quad \mathrm{k} = 1, \ldots, \mathrm{K} \tag{16.5b}$$

where k references all the observations in age cohort k, k′ is a specific state level age cohort, z is the intensity vector, y is output, x is input, θ is the inverse of the efficiency score, M is the number of outputs, N is the number of inputs, and K is the number of groups. The technology specified here is nonparametric, but assumes constant returns to scale and strong disposability of inputs and outputs.

The distance function specified in Eqs. (16.3) requires data from all the age cohorts and is computed for each observation k′ as:

$$\left(G_o^k\left(x^G, y^G\right)\right) = \max \theta^{k'} \tag{16.6}$$

subject to

$$\begin{aligned} &\sum_{k=1}^{K} z^k y_m^k \geq \theta^{k'} y_m^G \qquad m = 1, \ldots, M \\ &\qquad\qquad\qquad\qquad\quad \mathrm{n} = 1, \ldots, \mathrm{N} \\ &\sum_{k=1}^{K} z^k x_n^k \leq x_n^G \qquad \mathrm{k} = 1, \ldots, \mathrm{K} \\ &z^k \geq 0 \end{aligned}$$

In linear program models (16.5) and (16.6) each member of the z vector is bounded below by zero imposing constant returns. To impose variable returns, the constraint $\sum_{k=1}^{K} z^k = 1$ is added.

Because farm size varies, it is informative to ascertain the role of scale in productivity differences. Productivity consists of technology and technical efficiency. Technology is determined by dividing Malmquist computed from the Global technology, by the technical efficiency determined for each observation in

an age cohort using the technology used by that age cohort. Because total technical efficiency equals scale efficiency multiplied by pure efficiency, it is possible to decompose technical efficiency into scale and pure efficiency components by imposing that the z variables sum to one in distance function equation (16.5) and estimating the distance function under variable returns to scale. Then to determine the portion of technical inefficiency that is due to returns to scale, technical efficiency estimated under constant returns to scale is divided by efficiency estimated assuming variable returns to scale. The remaining technical efficiency is classified as pure efficiency.[2]

16.3 Results

16.3.1 Data

Data are obtained from the 2012 U.S. Agricultural Census, which is a complete enumeration of agricultural production entities in the United States. Individual farm observations are not publicly available, but data are summarized by age cohorts at the state level. The age cohorts are (1) under the age of 25 years, (2) age 25–34 years, (3) age 35–44 years, (4) age 45–54 years, (5) age 55–64 years, and (6) age 65 and older. Although some of the farms are owned by multiple individuals, organized into partnerships or corporate legal entities, the age of the principal operator is used to place a farm into an age cohort. The assumption is that the principal operator makes most or all final decisions. Burton (2006) makes the case that an index compiled by averaging the age of family members working on the farm would be better to study the life cycle phenomenon; those data are not published in the cohort groupings. Data are not available separately for sole proprietorships. Farms owned and managed by multiple individuals are almost exclusively family businesses, many parent-child operations.

Data are summarized from farms, where the respondent states that farming is his full time occupation, and separately into farms where the respondent states that farming is not his full time occupation. Only the data from farmers who indicated that farming is their full time occupation are used. Data are constructed for an average farm in each age cohort by dividing aggregate output or expense for the state for a specific age cohort by the number of farms in that age cohort in that state. Aggregate county level data are also available, but unfortunately those data are not summarized by age cohort, precluding using county level observations. Data are extracted from the web.

[2]The software used was Paul Wilson's FEAR program in the language R. The routine "dea" from FEAR was used to solve the various linear programs using an output orientation under constant and variable returns to scale. The linear programming solutions were then used to derive the various Malmquist and component indices.

One output is defined as total agricultural sales, which includes sales of all crops and animal products from the farm. Added to sales as the output are government agricultural payments received. This is done under the assumption that these payments occur because of farming activities and thus should be included as agricultural output.

Five inputs are defined: (1) crop input, (2) livestock input, (3) labor input, (4) operator labor input, and (5) capital input. Unfortunately, these aggregated categories are not how expenses are reported in the Census. Rather, more detailed itemized expense items are reported and summed into these five inputs. Items such as fertilizer, chemicals, seed, etc., are placed into crop inputs, while purchased feed, utilities, supplies, etc., are placed into livestock input. The detailed expenses from the Census and where they are placed is reported in Table 16.1. A simple arithmetic aggregation is used with the aggregate divided by the number of farms in an age-state cohort to arrive at per farm data. Non-disclosure rules are effective for some of the detailed expense items for some age cohorts in specific states so as to not disclose the data from specific farming operations. Those expense items are not available for summation leading to an incomplete measure of aggregate costs. When this occurs, that age cohort in that state is excluded from the analysis. This

Table 16.1 Receipt and expense items merged to produce the output and five inputs

Aggregated item	Disaggregated items
Agricultural output	Agricultural sales
	Government payments
Crop input	Seed expenses
	Fuel expenses
	Chemical expenses
	Custom work expenses
	Fertilizer expenses
Livestock input	Utility expenses
	Livestock purchased
	Feed expenses
	Supplies
	Miscellaneous expenses
Labor input	Hired labor expense
	Contract labor
Operator labor input	Calculated as 250 days available minus:
	25 days if response was 1–49 days off-farm work
	75 days if response was 50–99 days off-farm work
	150 days if response was 100–199 days off-farm work
	225 days if response was 200 plus days off-farm work
Capital input	Machinery lease payment
	Depreciation
	0.05 × machinery value
	0.05 × real estate value

often happens in the states with few farmers in the younger cohort, under age 25. Missing because of non-disclosure is often the hired labor or contract labor expense item, but other expense items such as custom work are sometimes the culprit.

Table 16.2 provides the number of state observations and sales by age cohort. Older farmers are well represented; data for younger farmers are often missing due to non-disclosure rules. Total aggregate sales peak for cohort 55–65. Although sales for the oldest cohort (over 65) are lower than the peak, they are still higher than the sum of sales for the three youngest cohorts under age 45. The productivity of older farmers matters.

The Census questionnaire asks for products sold and items purchased during the year 2012, rather than output produced and inputs used. For individual operations, there can be significant differences between production and sales or usage and expenses because of inventory changes. However, on average these variations should average out when we divide total sales or expenses by the number of farms to arrive at the sales and inputs used by the average farm in an age cohort for each state. Some of the expense measures are also stocks which are converted into flow measures. Machinery lease payments and machinery depreciation are used as the flow of machinery input, although lease payments or tax deprecation may differ from use flow or economic depreciation. The opportunity cost of machinery and real estate, computed as 5% of the machinery and real estate value, is added as a capital expense.

Unfortunately, the quantity of family labor is not recorded unless the family member is paid a wage, in which case family labor is included as hired labor expense. More problematic is that the Census questionnaire does not collect hours each operator worked on the farm, but rather asks the number of days each year that the operator worked off the farm. This work off the farm could involve tasks that range from serving as a director of a cooperative to driving a school bus. So from an assumed 250 available days of work, the mid-point of the days worked off the farm

Table 16.2 Number of state observations for which complete data were available and sales by age cohort

Age	Number of states (out of 50)	Sales (millions) (all 50 states)
Under 25	28	1.004
25–34	27	15.593
35–44	43	44.573
45–54	47	98.724
55–65	48	111.707
Over 65	47	77.722

Year 2012 U.S. Agricultural Census (principal operators)

The 27 states that comprise the balanced panel that is used in the empirical work consist of Alabama, Arkansas, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Massachusetts, Minnesota, Mississippi, Missouri, Nebraska, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, South Dakota, Tennessee, Texas, Virginia, Washington and Wyoming

Table 16.3 Descriptive statistics of inputs and output (sample = 27) mean and (standard deviation)

Age cohort	Crop expense	Livestock expense	Labor expense	Capital expense	Operator days	Sales
Under 25	41.6 (22.6)	61.7 (54.2)	5.3 (3.0)	62.8 (25.0)	210 (22.8)	161.8 (80.5)
25–34	73.8 (44.0)	111.0 (56.0)	14.1 (8.6)	99.7 (46.8)	213.3 (15.6)	301.5 (137.0)
35–44	105.0 (58.4)	215.7 (120.6)	29.3 (22.7)	144.3 (73.4)	221.7 (12.1)	509.1 (208.7)
45–54	108.0 (63.3)	203.6 (137.0)	31.1 (26.6)	153.0 (82.0)	229.6 (11.5)	500.5 (238.5)
55–64	92.0 (54.4)	145.8 (86.1)	25.7 (20.4)	134.8 (74.2)	247.1 (6.3)	389.6 (194.0)
Over 65	40.1 (27.3)	71.2 (48.4)	13.3 (10.5)	82.5 (41.4)	269.9 (4.19)	192.6 (105.6)

Note crop, livestock, labor, capital and sales are in thousands of dollars

interval is subtracted as shown in Table 16.1. Descriptive statistics for the core sample of 27 states are contained in Table 16.3.

It is useful to step back and summarize our procedure. The unit of analysis is a farmer age cohort in a state. Because of Census non-disclosure rules, we have an unbalanced panel of states over 6 age cohorts as shown in Table 16.2. We pool these data and calculate the global constant returns to scale frontier based upon these 240 observations. To determine productivity change (Malmquist) we estimate the distance of each of the 240 observations relative to the global frontier. This provides the Malmquist for each of the 240 observation. Then to determine technical efficiency for an observation in an age cohort, we evaluate that observation relative to a frontier based only upon data for all state observations from the same age cohort. Technical efficiency is then decomposed into scale efficiency and pure efficiency given the technology used by an age cohort. Thus technical efficiency of the farmers under the age of 25, for instance, is determined relative to only the other farmers under the age of 25. Technical change is calculated as Malmquist productivity from the global technology divided by efficiency determined within an age cohort. Because the 25–34 age cohort has the fewest number of complete state observations at 27, we construct indices for only those 27 states as identified in Table 16.2. This allows matching state results across all age cohorts.

16.3.2 Results

Malmquist productivity is calculated with total sales of agricultural products as the output and five inputs: crop expenses, livestock expenses, operator labor in days, other labor expenses, and capital expenses. The year is 2012. Cross section Malmquist substitutes the six age cohorts for time periods. All variables are state averages for each age cohort. Efficiency, technology, scale technology and productivity indices are calculated using linear programs (5) and (6) at the state level, using the FEAR software package authored by Paul Wilson (2008).

Although all available data were used to compute distance functions, in order to derive consistent indices for making comparisons across all age cohorts, we construct a balanced panel from the computed distance functions that is driven by age cohort 25–34, which has complete data for 27 states. These states are listed in Table 16.2. Data are available for these 27 states for all other age cohorts. The major agricultural states are included except for California because of disclosure rules for the youngest age group in that state, although other age groups from California for which data were available defined the technology sets. In order to clearly make comparisons across age cohorts, we normalize the results to the youngest age cohort so the index value for the youngest age cohort is one. This comparison makes all of the indices transitive, which is generally not the case for Malmquist indices using adjacent ages (Førsund 2002).

As a point of comparison, Tauer and Lordkipanidze (2000) present results for efficiency, technology and productivity indices for the 1992 Census aggregated into regions. The regional results from the 1992 Census are reasonably consistent. Efficiency modestly rises to age cohort 35–44 and is then flat for the remaining age cohorts. Technology generally rises, peaks at age cohort 35–44 and then falls. Productivity mimics the results for technology. The Tauer and Lordkipanidze (2000) study applies conventional Malmquist to adjacent age cohorts rather than the global approach across all age groups used here.

Our results using data 20 years later are different. Consider the 2012 results shown in Fig. 16.2, which are presented like traditional Malmquist indices except that age cohorts rather than time periods are on the horizontal axes. Age cohort data are the geometric means of state results and represent percentage deviations from the youngest cohort group. The values that underlie the figures are contained in Table 16.4. Efficiency relative to the youngest cohort rises, dips slightly, rises through age cohort 55–65 and then falls to age cohort 65 and older. However, since our focus is on what happens to efficiency as farmers get older, it is interesting that the most efficient farmers are 55–65, who are 7% more efficient than the youngest

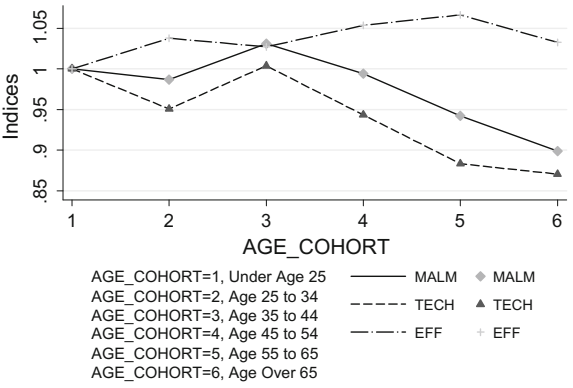


Fig. 16.2 Malmquist, efficiency and technology differences by age cohort of 2012 U.S. Agricultural Census Farmers. *Data Source* USDA NASS, Census of Agriculture

Table 16.4 Malmquist and its decompositions relative to the youngest age cohort

Age cohort	Malmquist	Technology	Efficiency	Scale efficiency	Pure efficiency
Age under 25	1.00	1.00	1.00	1.00	1.00
Age 25–34	0.99	0.95	1.04	0.95	1.09
Age 35–44	1.03	1.00	1.02	0.97	1.06
Age 45–54	0.99	0.94	1.05	0.97	1.08
Age 55–64	0.94	0.88	1.07	0.97	1.10
Age over 65	0.90	0.87	1.03	0.96	1.07

cohort, and the oldest cohort is 3% more efficient than the youngest cohort, and slightly more efficient than farmers 35–44. The experience of the oldest farmers enables them to maintain a position relatively close to the prevailing frontier, particularly compared to farmers entering the industry. This result is reasonably consistent with Tauer and Lorkipanidze (2000) who conclude that older farmers are largely able to maintain efficiency.

The technology results tell a different story. The youngest farmers adopt the latest technology at the onset of their agricultural careers; 5% more efficient in technology than the next oldest cohort, 25–34. Why this dip occurs is interesting and could be because of a number of reasons. This age group, if entered farming 10 years earlier as under 25 years of age farmers, may have faced financial constraints limiting their acquisition of state of the art technology. Alternatively they may simply have purchased technology ten years ago that is not as productive as current technology. Farmer cohort 35–44 regains a position on the technology frontier compared to the youngest, possibly because they may have cycled to new technology. Then the older age cohorts lose ground, concluding with farmers over 65 using 13% less technology than the youngest. There appears to be a fundamental difference between maintaining a position relatively close to a familiar frontier and keeping up with an expanding frontier that is driven by technology.

The Malmquist results mimic the technology results as the magnitudes of the technology comparisons are larger than the efficiency comparisons. Overall, the oldest farmers are 13% less productive than the most productive age cohort, 35–44, and 10% less productive than the youngest. From a policy perspective, this underscores the importance of policies that address the barriers to older farmers responding to technical change. Tauer and Lorkipanidze (2000) also conclude that the challenge to older farmers is driven by keeping up with technology.

Figure 16.3 decomposes efficiency into pure efficiency and scale efficiency to gain insights into to what extent farmer efficiency is driven by changes in size or changes in their position relative to the frontier holding size constant (Fare et al. 2008). It is important to remember that all efficiency measures were measured using only the reference technology set of the age cohort for whom efficiency was

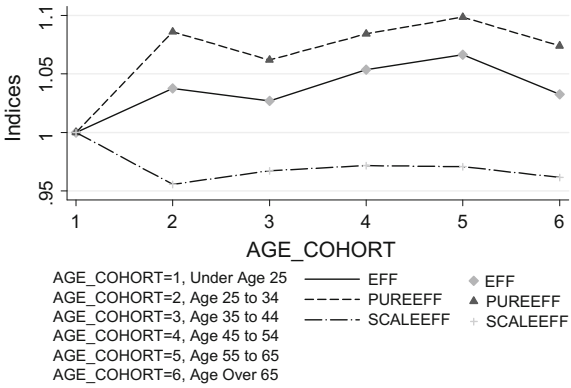


Fig. 16.3 Efficiency, pure efficiency and scale efficiency by age cohort of 2012 U.S. Agricultural Census Farmers

calculated. The global reference set was used to measure technology differences only. Average age pure efficiency across age is higher than efficiency generally by about four percentage points (relative to the youngest cohort) and exhibits the same pattern. Scale efficiency is generally around 8% lower than efficiency (relative to the youngest cohort) and exhibits a different pattern—it falls by 5% between the youngest and age cohort 25–34 and then remains fairly flat. Pure efficiency drives the pattern of efficiency across age cohorts and scale pulls efficiency down. Farms of sub-optimal size exert a negative impact on operating close to the existing frontier. This effect sets in immediately between the youngest farmers and the next age cohort and then remains constant across subsequent age cohorts. Older farmers do not behave differently than younger farmers in this case. This might have been expected since efficiency for each state observation in an age cohort is measured relative to the other farmers in that age cohort across all states.

The discussion up to this point has focused on the pattern of Malmquist and its components based upon the geometric means for each state age cohort. The aggregation by age cohort using the geometric mean hides the distribution of the results for each age cohort. Figures 16.4, 16.5 and 16.6 focus on the distribution of productivity and its decompositions for each age cohort. The tighter the distribution, the more confident we can be that the geometric mean is a good measure. The graphs are kernel densities of histograms. For the most part these distributions are similar but right shifted in agreement with their geometric means in Table 16.4 and Figs. 16.2 and 16.3. The distribution of age cohort 45–44 is comparatively tight for Malmquist and that is due to efficiency as shown in Fig. 16.5. The distribution for technology is also comparatively tight for age cohort 35–44. The technology in Fig. 16.6 shows some slight bimodal distribution for age cohort 55–65, and the technology distribution for the age cohort 35–44 is more right shifted compared to the other age groups. The dispersed distribution for the indices suggest that more is going on than is revealed by the geometric mean aggregations used in the previous discussions.

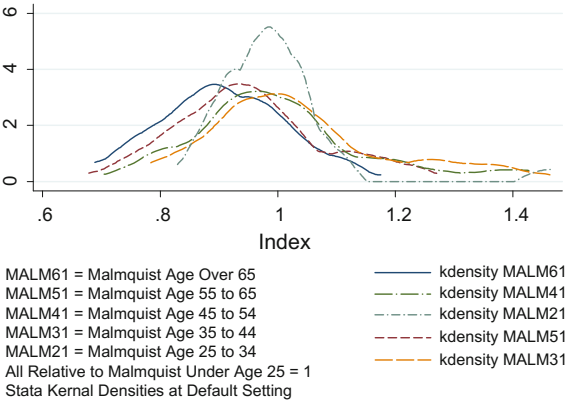


Fig. 16.4 Distribution of malmquist indices of the various U.S. farmer age cohorts

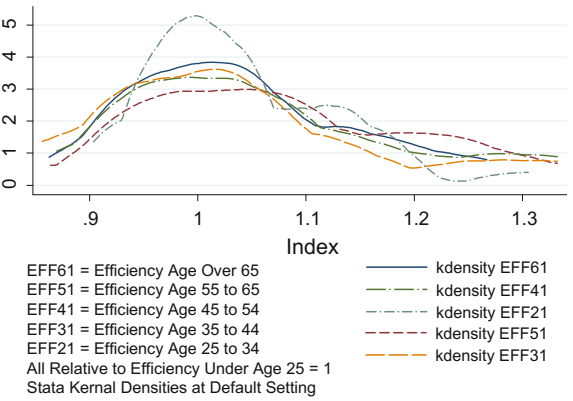


Fig. 16.5 Distribution of Efficiency Indices of the various U.S. farmer age cohorts

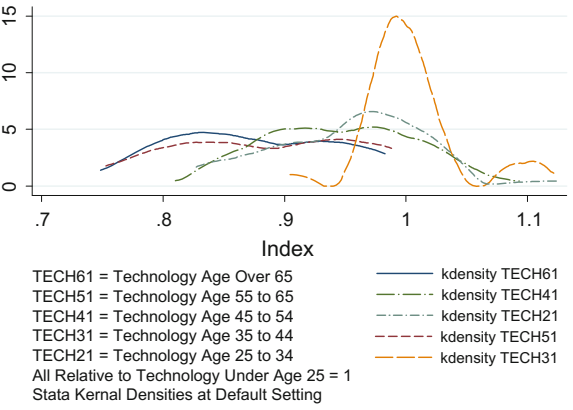


Fig. 16.6 Distribution of technology indices of the various U.S. farmer age cohorts

16.4 Conclusion

The benefit of applying a frontier approach to understanding the performance of farmers of different ages is that it produces measures of productivity, technology, efficiency, scale efficiency and pure efficiency that can be woven into a coherent story to shed light on the relationship of productivity and age and customize appropriate remedial policies. Overall, older farmers produce less given inputs than do younger farmers. Yes, we should worry about this. Farmers over 65 are 13% less productive than the most productive age cohort, 35–44 and 10% less productive than the youngest age cohort. Muting the impact of this lower productivity is that the older farmers have much lower sales (partially due to low productivity) than do farmers in the middle age cohort.

But what is driving this discrepancy in performance? Interestingly, it is not efficiency using their current technology. Older farmers on average are 6% more efficient than the youngest age cohort and lag behind the most efficient age cohort, 55–64, by only around 3%. Old farmers know how to stay close to a familiar frontier. The answer is technology. The youngest age cohort appear to be benefiting from technology at the outset of their careers while the oldest cohort lags behind by 12%.

Technology has changed the nature of farming, and like many other industries, the pace of technological change in agriculture has accelerated. The hours needed for previous multiple tillage operations to grow crops have been reduced as farming moved to minimum or no till in recent years. Corn is still planted using a tractor, but now the tractor with GPS steers itself down the field while the farmer can use his smart phone to check on commodity prices (or watch a major league baseball game). Harvesting equipment cabs are air conditioned and heated, resulting in more comfort. Backbreaking days of shoveling grain or pitching manure are mostly over.

Old farmers appear not to be fully benefiting from the technology revolution in agriculture. Since we know that farmers are getting older as fewer young people enter the profession, this takes a toll on agricultural productivity. Government policies to provide technical support would help to address this problem.

This paper also makes methodological contributions. It revives the technique to apply Malmquist to a cross section. This requires a variable that substitutes for time; in our case, the variable is age cohorts. An advantage of the cross section approach is that it eliminates noise that is introduced by the passage of time, particularly the vagaries of the weather, although it introduces noise due to the movement from one age cohort to another. It also revives a more recent development to use a global frontier as a fixed benchmark.

References

Asmild M (2015) Frontier differences and the global Malmquist index. In: Zhu J (ed) Data envelopment analysis handbook of models and methods. Springer, Boston, pp 447–462

Burton RJF (2006) An alternative to farmer age as an indicator of life-cycle stage: the case for a farm family age index. J Rural Stud 22:485–492

Camanho AS, Dyson RG (2006) Data envelopment analysis and Malmquist indices for measuring group performance. J Prod Anal 26:35–49
Fairweather J, Mulet-Marquis S (2009) Changes in the age of New Zealand farmers: problems for the future? NZ Geogr 65:118–125
Fare R, Grosskopf S, Margaritis D (2008) Efficiency and productivity. In: Fried HO, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency and productivity growth. Oxford University Press, New York, pp 522–622
Feyrer J (2007) Demographics and productivity. Rev Econ Stat 89(1):100–109
Førsund FR (2002) On the circularity of the Malmquist productivity index. International centre for economic research ICER working papers number 29-2002
Fried HO, Tauer LW (2011) The impact of age on the ability to perform under pressure: golfers on the PGA tour. J Prod Anal 35(1):75–84
Fried HO, Tauer LW. (2012) Age and performance under pressure: golfers on the LPGA tour. In: Shmanske, Stephan, Kahane, Leo H. The oxford handbook of sports economics, volume 2 economics through sports. New York, Oxford University Press: 135–152
Gale HF (1994) Longitudinal analysis of sarm Size o\over the farmer's life cycle. Rev Agric Econ 16(1):113–123
Katchova AL, Ahearn MC (2015) Dynamics of farmland ownership and leasing: implications for young and beginning farmers. Appl Econ Perspect Policy 1–17
Kurtzleben D (2014) The rapidly aging U.S. farmer. Data Mine, www.usnews.com/news/blogs/data-mine/2014/02/24/us-farmers-are-old-and-getting-much-older, accessed Oct 2015
Mishra AK, El-Osta HS (2008) Effect of agricultural policy on succession decisions of farm household. Rev Econ Household 6:285–307
Oster SM, Hamermesh DS (1998) Aging and productivity among economists. Rev Econ Stat 80 (1):154–156
Pastor JT, Lovell CAK (2005) A global Malmquist productivity index. Econ Lett 88:266–271
Skirbekk V (2003) Age and individual productivity: a literature survey. Max-Planck institute for demographic research working paper WP 2003-028
Tauer LW (1984) Productivity of farmers at various ages. North Central J Agric Econ 6(1):81–87
Tauer LW (1995) age and farmer productivity. Rev Agric Econ 17(1):63–69
Tauer LW, Lordkipanidze N (2000) Farmer efficiency and technology use with age. Agric Resour Econ Rev 29(1):24–31
Vilsack T Agricultural secretary Vilsack announces grants to support beginning farmers and ranchers across 24 states, (http://www.usda.gov/wps/portal/usda/usdahome?contentidonly=true&contentid=2012/08/0287.xml), accessed 28 Oct 2015
Wilson PW (2008) FEAR 1.0: a software package for frontier efficiency analysis R. Socio-Economic Plann Sci 42:247–254

Author Index

J. Aparicio et al. (eds.), *Advances in Efficiency and Productivity*,
International Series in Operations Research & Management Science 249,
DOI 10.1007/978-3-319-48461-7