The Palgrave Handbook of Economic Performance Analysis

Edited by
Thijs ten Raa · William H. Greene

palgrave
macmillan

The Palgrave Handbook of Economic Performance Analysis

Thijs ten Raa · William H. Greene
Editors

The Palgrave Handbook of Economic Performance Analysis

palgrave
macmillan

Editors
Thijs ten Raa
Utrecht School of Economics
Utrecht University
Utrecht, The Netherlands

William H. Greene
Stern School of Business
New York University
New York, NY, USA

ISBN 978-3-030-23726-4 ISBN 978-3-030-23727-1 (eBook)
https://doi.org/10.1007/978-3-030-23727-1

© The Editor(s) (if applicable) and The Author(s) 2019
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cover illustration: Anna Berkut/Alamy Stock Photo
Cover design by eStudio Calamar

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

List of Figures

Microfoundations of Earnings Differences

The Choice of Comparable DMUs and Environmental Variables

Data Envelopment Analysis with Alternative Returns to Scale

Ranking Methods Within Data Envelopment Analysis

Nonparametric Statistical Analysis of Production

Bayesian Performance Evaluation

Common Methodological Choices in Nonparametric and Parametric Analyses of Firms' Performance

Modelling Environmental Adjustments of Production Technologies: A Literature Review

An Overview of Issues in Measuring the Performance of National Economies

Productivity Indexes and National Statistics: Theory, Methods and Challenges

List of Tables

Microfoundations of Earnings Differences

The Choice of Comparable DMUs and Environmental Variables

Ranking Methods Within Data Envelopment Analysis

Stochastic Frontier Models for Discrete Output Variables

Nonparametric Statistical Analysis of Production

Bayesian Performance Evaluation

Intermediate Inputs and Industry Studies: Input-Output Analysis

An Overview of Issues in Measuring the Performance of National Economies

Introduction

William H. Greene and Thijs ten Raa

During the June 2015 European Workshop on Efficiency and Productivity Analysis in Helsinki, we sat together and tried to map a broad view of performance analysis. The prevailing approach is frontier analysis. The production frontier of a "decision making unit" (such as a firm, an industry, an economy, or conglomerates of the aforementioned) maps the maximum amount of output given the available input, where input and output are multidimensional objects, comprising different types of labor, capital, intermediate inputs, goods, services, and other products. If the "distance" between the actual input-output combination of a decision making unit and the frontier is small, then the unit is efficient. If the frontier is far out, then efficient units are productive. This framework is amenable to precise measurements of efficiency and productivity, but numerous issues surround it and cast shadows on numerical results. This compendium collects a set of works that explore these issues.

First, the framework suggests there are given, fixed lists of input components and output components, but what about new inputs, outputs, and

W. H. Greene
Stern School of Business, New York University, New York, NY, USA
e-mail: wgreene@stern.nyu.edu

T. ten Raa (✉)
Utrecht School of Economics,
Utrecht University, Utrecht, The Netherlands
e-mail: tenraa@uvt.nl

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_1

intermediate products? Innovation ought to play a role in performance analysis. Second, the production frontiers reflect the best practices, or benchmarks, of producing outputs given inputs. But are these benchmarks relevant to competing decision making units? Observed or unobserved conditions may be favorable to the benchmarks. Third, in which direction do we measure the distance between an actual input-output combination and the frontier? The literature analyzes alternatives and shows when results differ and when they don't. But the bulk of it is mechanical and we believe this issue should be related to the objective of an economy. Fourth, what is the appropriate or most informative level of analysis, individuals, firms, industries, or economies? The majority of frontier analyses are industry studies, such as banking, and the same techniques, particularly data envelopment analysis and stochastic frontier analysis, are applied without much further ado at either a more microlevel, such as productivity rankings of workers, or a more macrolevel, such as the performance of national economies. However, the more aggregated the level of analysis, the more scope there is for eliminating sources of inefficiencies, because of the assumptions that inputs are considered given water down, or at least are averaged away. Fifth, by directly working with inputs and outputs, frontier analysis can examine the role that prices pay. Technical frontier analysis does not need price information. Yet prices play a shadow role. In evaluating the frontier output given the available inputs, one solves a mathematical program, and the Lagrange multipliers of the constraints are shadow prices which measure the productivities of the inputs. This raises the issue which prices are relevant: observed market prices or shadow prices?

These issues are the focus of this Handbook. To address them, we will review new developments in frontier analysis. We will extend the analysis by including contributions to the performance literature which we think throw fresh light on the issues. A topic *not* included in the Handbook was engineering performance, such as reliability analysis. The title of the volume is *Handbook of Economic Performance Analysis*. Economic theory is also used to organize the related issues of distance directions, objective functions, and performance measures.

Although the focus of the performance literature is on production, we believe that the issues can also be addressed by bearing in mind that the ultimate aim of an economy is to serve the well being of consumers. Consumers maximize utility $U(x)$ subject to the budget constraint $px \leq p\omega$, where x is an n-dimensional commodity bundle, p is the n-dimensional market price (row) vector, and ω is the n-dimensional endowment of the consumer. The first-order condition is that the marginal utilities are proportional to the

prices, $\partial U/\partial x = \lambda p$, where $\lambda \geq 0$ is the marginal utility of income. Hence, the direction of steepest utility increase is the market price, irrespective of the specifics of utility, i.e., for all consumers. So market price is a natural direction to measure efficiency. Productivity, however, measures the level of utility attainable given the resources. Utility differences do not matter, so the principle is easiest explained by assuming all consumers have the same utility function U. The frontier is determined by maximizing $U(x)$ subject to the material balance constraint $x \leq y + \omega$ where y is an n-dimensional member of the production possibility set Y and ω is the n-dimensional endowment of all consumers. This maximization problem does feature prices. Assuming constant returns to scale in production (at the aggregate level), the first-order condition is that the marginal utilities are proportional to the prices, $\partial U/\partial x = p^*$, where $p^* \geq 0$ is now the shadow price vector of the balance constraint. So shadow price is a natural direction to measure productivity.

Performance is related to both efficiency and productivity. Roughly speaking, performance can be improved by raising efficiency or by pushing out the frontier. This Handbook presents useful formulas that illuminate the connections in a variety of settings. Clearly, an important, complicating issue is the choice of price. The Handbook discusses more such issues and we think that this broad approach will prove fruitful in advancing performance analysis. For example, the markup between a market price and the shadow price measures market power and this insight may be used to interconnect the efficiency and productivity components of performance in an industrial organization framework.

We attempt to bridge the gap between the two main methodologies of performance analysis, data envelopment analysis (nonparametric) and stochastic frontier analysis (parametric). Often it is not clear what variables are included, if they can be controlled, what their nature is, from a mathematical or statistical point of view. For example, do some firms perform better because their technology is superior or because they are more innovative? Standard practice is to divide variables between inputs, outputs, and environmental variables, to use the inputs and outputs for performance measurement using either methodology, and finally to analyze the relationship between performance results and the environment. This practice, however, raises theoretical and statistical issues. We approach the methodological issues by reviewing variants of stochastic frontier analysis, e.g., with discrete variables and alternative distributions, and by exploring statistical analyses of nonparametric performance measurement, by Bayesian analysis. We will conclude by reviewing commonalities between nonparametric and parametric analyses.

One of the novelties of this Handbook is the coverage and interrelation of microeconomic, mesoeconomic, and macroeconomic performance analysis, in this order, intertwined by methodological contributions.

The first topic is "Micro Foundations of Earnings Differences," in next chapter, by Tirthatanmoy Das and Solomon W. Polachek. The authors review the distribution of earning differences. The paper finds one's ability to learn and one's ability to retain knowledge are most influential in explaining earnings variations. The chapter is a detailed overview of the human capital view of earnings differentials.

In chapter "Performance: The Output/Input Ratio," Thijs ten Raa critically reviews and interrelates the fundamental concepts of performance analysis, including total factor productivity, the Solow residual, the Farrell efficiency measure, the Debreu-Diewert coefficient of resource utilization, and the Malmquist, Törnqvist, and Fisher indices, for alternative numbers of outputs and inputs.

The main drivers of performance analysis are reviewed in chapter "R&D, Innovation and Productivity," by Pierre Mohnen. He analyzes the indicators used to perform firm R&D, innovation, and performance analyses and explains the theoretical link between innovation and productivity growth. He then considers the estimated magnitudes in that relationship using the different innovation indicators.

In chapter "The Choice of Comparable DMUs and Environmental Variables," John Ruggiero addresses the important issues of choosing comparable decision making units and environmental variables in efficiency analysis. In standard nonparametric analysis, decision making units produce common outputs from common inputs, under common conditions. There are methodological trade-offs. On the one hand, the commonality assumptions are better fulfilled by limiting the analysis to smaller numbers of more comparable decision making units. As a consequence, however, the reduction in the number of potentially competing benchmarks increases the efficiency estimates, often toward 100% when the number of production units goes down to the number of outputs. And if the number of environmental values goes up to the number of production units, all inefficiencies may be explained away as well.

Scale efficiency is an important form of efficiency. Shubash Ray analyzes this case in chapter "Data Envelopment Analysis with Alternative Returns to Scale." Production units are more efficient when they operate at a scale with lower average costs and data envelopment analysis is a convenient methodology in which to model this outcome.

In chapter "Ranking Methods Within Data Envelopment Analysis," Nicole Adler and Nicola Volta present the issue of rankings. They address the lack of discrimination in DEA applications, in particular when the number of inputs and outputs is relatively high in comparison with the number of production units, borrowing techniques from the social sciences, including multiple-criteria decision making.

The Handbook proceeds in chapter "Distributional Forms in Stochastic Frontier Analysis," to stochastic frontier analysis. In chapter "Distributional Forms in Stochastic Frontier Analysis," Alexander Stead, Phill Wheat, and William Greene survey the developments in stochastic frontier modeling. The basic function in this literature is models of errors of measurement based on the normal distribution and (in)efficiency, a signed concept, by the half-normal distribution. The former distribution is well vested in theory, building around the central limit theorem. But the latter is ad hoc and various alternatives have emerged. These are reviewed from a practitioner's point of view.

In chapter "Stochastic Frontier Models for Discrete Output Variables," Eduardo Fé addresses another important issue, encountered in labor, industrial, and health economics, where outputs are nontangible and nonpecuniary outcomes and often measured through indicators of achievement (employment status, academic certification, success in a labor market scheme), ordered categories (like scales describing job satisfaction, health status, personality traits), or counts (numbers of patents or infant deaths). This chapter generalizes the standard stochastic frontier model to encompass such situations.

The gap between the two main frontier methodologies has attracted developers of intermediate approaches, which are reviewed in the next three chapters. Chapter "Nonparametric Statistical Analysis of Production," by Camilla Mastromarco, Leopold Simar, and Paul Wilson, analyzes the data envelopment analysis of chapters "The Choice of Comparable DMUs and Environmental Variables", "Data Envelopment Analysis with Alternative Returns to Scale" and "Ranking Methods Within Data Envelopment Analysis," but attaches that approach to a statistical analysis. Their approach requires large samples of, e.g., production units (with their input-output combinations), but then combines the strengths of nonparametric analysis and stochastic frontier analysis to make statements about expected performance measures and their confidence intervals.

Ultimately distributional assumptions differentiate the different approaches presented thus far. A natural variant is reviewed in

chapter "Bayesian Performance Evaluation," by Mike Tsionas. The Bayesian approach provides a formal and straightforward way to facilitate statistical inference, which is always a problem in data envelopment analysis, despite the recent advances exposited in the previous chapters.

Chapter "Common Methodological Choices in Parametric and Nonparametric Analyses of Firms' Performance," by Luis Orea and José Zofío finds common ground for the two frontier analyses, using distance functions as the organizing principle, which is standard in data envelopment analysis, but novel in stochastic frontier analysis. Key issues related to alternative technological assumptions and alternative economic objectives of the firm are thus reviewed. The issues of the number of production units relative to the numbers of outputs and inputs and the number of environmental variables are addressed. The choice of direction of the distance function is discussed.

The next four chapters take us from microproduction analysis to meso- and macroeconomics. The crucial choice of prices is related to the objective of an economy in chapter "Pricing Inputs and Outputs: Market Prices Versus Shadow Prices, Market Power and Welfare Analysis," by Aditi Bhattacharyya, Levent Kutlu, and Robin C. Sickles. Market prices cannot be simply taken to signal the social value of production. There are two ways forward: (i) correct the observed market prices to derive the shadow price and (ii) try to derive the shadow price directly without using market prices (e.g., when market prices do not exist). The chapter surveys the methods used in the literature to derive social valuations when market prices are not accurate or not available due to the following considerations: imperfect competition, effects on income distribution and growth that are not factored into prices and external effects, like on the environment.

Chapter "Aggregation of Individual Efficiency Measures and Productivity Indices," by Andreas Mayer and Valentin Zelenyuk, reviews the key existing results on aggregate efficiency measures and aggregate productivity indices and outlines new results for the aggregation of the Hicks-Moorsteen productivity index, and outlines some insights into ongoing and future directions of research in this area.

Chapter "Intermediate Inputs and Industry Studies: Input-Output Analysis," by Victoria Shestalova, reviews performance measurement for industries and the whole economy, taking into account the interindustry deliveries, using the workhorse of applied general equilibrium analysis, the modern input-output model with possibly different numbers of inputs, outputs, and industries. The methodology, a synthesis of frontier

and input-output analyses, is suitable for performance measurements within both national and international industrial studies, environmental analysis, and other policy-relevant analyses. Data requirements and international databases are reviewed, as are applications, including the assessment of emission caps and pricing policies supporting the adjustments in industrial outputs.

In chapter "Modeling Environmental Adjustments of Production Technologies: A Literature Review," Hervé Dakpo and Frederic Angwe present a theoretical discussion of negative externalities, including the production of "bads," without an explicit allusion to performance. They summarize the lessons from the different models in the literature and the challenges that need to be dealt with in modeling environmentally adjusted production technologies.

The last two chapters are macroeconomic. In chapter "An Overview of Issues in Measuring the Performance of National Economies," Anthony Glass, Karligash Kenjegalieva, Robin Sickles, and Thomas Weyman-Jones measure the aggregate economic performance of national economies, considering a wide range of different measures including the value-added definition of GDP and economic welfare. They show how stochastic frontier analysis and data envelopment analysis modeling has been able through the idea of total factor productivity (TFP) decomposition and the measurement of inefficiency to tell us much more about TFP than the more conventional approaches. They review the issue of whether the performance of national economies converges over time, or whether, as suggested by endogenous growth models, the individual performance of different countries is endogenous to the country itself. Technological spillovers among neighboring countries at the level of the aggregate production function are analyzed.

In chapter "Productivity Indexes and National Statistics: Theories, Methods and Challenges," Erwin Diewert and Kevin Fox provide the theoretical justifications for the index number formulae for productivity growth measurement that are commonly used by national statistical offices. They then turn to a discussion of data used in index number construction in practice and highlight the measurement challenges. The choice of index number formula is examined based on an "axiomatic" approach and from the perspective of economic theory, recognizing that the resulting indexes are measuring economic concepts. The results provide the justification for the index number choices made by national statistical offices in constructing productivity growth estimates. Data needs for constructing the productivity indexes are discussed and the concepts, sources, and methods that are used for the

output, labor, and capital components are reviewed. National statistical offices face several difficult measurement problems and this chapter suggests ways forward.

We are grateful that our proposal to edit a Handbook which highlights issues of mainstream efficiency and productivity analysis and offers a broad perspective on economic performance analysis, was well received and accepted by the chapter contributors we had in mind and thank them for their work. We are also grateful to the referees who wrote anonymous reports, but whom we now reveal. Thank you Antonio Amores, Bert Balk, Jan Boone, Walter Briec, Maria da Conceição Andrade e Silva, Rolf Färe, Shawna Grosskopf, Reza Hajargasht, Joop Hartog, Jens Krüger, Chris O'Donnell, Raquel Ortega-Argilés, Inmaculada Sirvent Quilez, Mark Steel, and Emmanuel Thanassoulis.

Microfoundations of Earnings Differences

Tirthatanmoy Das and Solomon W. Polachek

1 Introduction

Modern labor economics considers workers as a conglomeration of heterogeneous units each differing in productivity. As such, most labor economists now focus on how skills differ between individuals, and as a result, how dissimilar capabilities give rise to each worker commanding a different wage. Thus, rather than concentrating on the functional distribution of income between labor and capital, as had been the case in the past, economists now focus more attention to pay differences across various segments of the population. Indeed, some of these differences have vastly widened in the last

T. Das (✉)
Economics and Social Sciences Area, Indian Institute of Management Bangalore, Bangalore, India
e-mail: tirthatanmoy.das@iimb.ac.in

T. Das · S. W. Polachek
IZA Institute of Labor Economics, Bonn, Germany

S. W. Polachek
Department of Economic, State University of New York at Binghamton, Binghamton, NY, USA
e-mail: polachek@binghamton.edu

Liverpool Hope University, Liverpool Hope, UK

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_2

35 years.[1] This chapter examines the microeconomic basis of such variations in earnings, why they occur, and why they have changed over time.

We begin by examining patterns in current data. Repeatedly and overwhelmingly, one finds earnings are significantly correlated with one's years of school and one's age. Indeed, education appears to be the surest path to success, as all data indicate an individual's earnings to be higher the greater the years of schooling completed. With regard to age, one typically observes earnings to rise as one gets older, but at a diminishing rate. Earnings also vary by occupation, industry, size of firm, location, and a myriad of other factors. But there are other patterns too: Males earn more than females, whites earn more than blacks, but the gender gap within race is smaller for blacks than whites. Single childless women earn almost as much as single men, but married women lag far behind married men. Children exacerbate the gender wage gap. Immigrants earn less than natives, but over time in the receiving country, immigrant earnings eventually converge to natives' earnings.

Many theories have been used to explain *some* but not all these patterns. These include stochastic models entailing sheer luck, whereby circumstances largely outside one's control determine success; agency models whereby wage structures perhaps instigated by institutional forces such as tax policy or unions determine well-being; efficiency wage models that link wages to unemployment; matching models which account for why job turnover declines with tenure; crowding models that elucidate why women earn less than men; screening models which describe why education enhances earnings; occupational segregation models that portray why women are in lower-paying occupations than men; and productivity-enhancing contract models that provide an explanation for upward sloping age-earnings profiles. Whereas each of these theories has some predictive power, they individually deal with a single narrow aspect of earnings. In our opinion, only the life-cycle human capital model appears to explain the preponderance of all patterns simultaneously. Thus, this chapter focuses on human capital theory and the empirical work emanating from it.

Human capital theory postulates a person's earnings capacity to be directly proportional to his or her labor market skills and knowledge,

[1]One reason for the increased attention, at least in the USA, stems from the rising share going to labor until the 1970s (Krueger 1999; Armenter 2015). However, of late, there has been a reversal of this trend and a renewed interest in the functional distribution of income, only now dealing with the rising share to capital, especially since the 2000s (Mike Elsby et al. 2013; Karabarbounis and Neiman 2013). One new theory attributes this change to firm heterogeneity. In particular, Autor et al. (2017) describe "superstar firms" where labor's share fell relatively more.

collectively known as human capital. Each year a person augments human capital stock by the amount of new human capital he or she creates, and diminishes it by the amount he or she depreciates. Creating new human capital entails combining time and existing human capital. The greater one's ability, the more human capital one can produce, and the more rapidly one's earnings rise.

Of course, measuring ability is tricky. Most studies use IQ or achievement tests, but these standardized tests have been criticized because they get at analytic academic capabilities that usually lead to success in school, but not necessarily a proficiency that translates into real-world accomplishments (Sternberg 1985). On the other hand, the human capital model contains five parameters related to the production of human capital. Of these, three correspond directly to one's ability to create human capital, one to skill depreciation, and one to a person's time discount rate. New research (Polachek et al. 2015) enables one to back out these parameters for individuals rather than the population as a whole and show how they relate to labor market success.

These human capital parameters are also important in other domains of economics. For example, they are used in earnings dynamics models (Blundell 2014; Hoffmann 2016; Meghir and Pistaferri 2011), dynamic general equilibrium models (King and Rebelo 1999), but more importantly they are used to interpret skill formation (Cunha et al. 2006) in understanding earnings distributions. Typically, due to the lack of panel data and cumbersome computation, past studies estimate these parameters population-wide in a more or less representative agent framework. However, representative agent models are limited and can yield misleading inferences (Browning et al. 1999). Polachek et al. (2015) estimate these parameters person by person. Getting at these individual-specific human capital parameters enables them to test predictions of the human capital model. It allows us in this chapter to evaluate the importance of ability and other factors in shaping the earnings distribution. Our results suggest these five measures to be the most important explanatory factors related to labor earnings.

Much current research adopts a Mincer's log-linear specification of the human capital model. Many implications emerge from such analyses. These include how earnings rise with age at a diminishing rate over the life cycle, how earnings differ by demographic group, but most important how school relates to earnings. Early studies viewed education as an exogenous variable and obtained estimates of rates of return to schooling. However, because many recognized the endogeneity of schooling, subsequent researchers utilized quasi-experimental analyses to assess the value of education.

Quasi-experimental methods are not the panacea for identification. For example, instrumental variable results estimating the rate of return to schooling vary widely between 3.6 and 94.0% (Card 2001). Many question the validity of the exclusion restriction requirement for these instruments. But independent of this criticism, employing linear models, as these studies typically do, necessarily yields erroneous parameter estimates even with a valid instrument. This is because there exists an omitted nonlinear component of the earnings–schooling relationship which then constitutes a part of the error term. As such, the instrumental variable is correlated with the error. Moreover, ignoring heterogeneity further exacerbates the endogeneity problem.

We begin this chapter by describing inherent earnings patterns. We argue these patterns can be explained using the life-cycle human capital model. We utilize a simplified Mincer formulation and its extensions to explore observed demographic earnings differences and their trends. We then utilize the five parameters mentioned above which have been obtained from a structurally derived complex nonlinear earnings function and discuss their implications regarding predictions obtained from theory. From here, we exploit person-specific differences in these five parameters to explain earnings inequality. We concentrate on determining the importance of ability compared to the importance of schooling. We review studies that evaluate the impact of schooling using OLS and quasi-experimental approaches, and then we explain their pitfalls. We conclude by showing that the ability parameters obtained from structural human capital earnings function models are the most important determinants of earnings distribution. From a policy perspective, we claim treatments that enhance ability such as through early childhood interventions are the most effective in reducing earnings inequality.

2 Earnings Patterns

Earnings differ by age, schooling level, gender, race, and many more demographic factors. Understanding why these differences arise is important because the answers can help improve individual and societal well-being. Policymakers can use the answers to devise strategies to help ease poverty and eventually to help put countries on a path of increased growth and prosperity. To set the stage, we examine a number of these demographic earnings differences. We do so in five tables and one figure.[2] Each explores aspects of earnings inequality.

[2]These tables update data previously presented in Polachek (2008).

Table 1 depicts average US weekly wages in 2000 dollars by race, gender age, and education. As can be seen, women earn less than men, and blacks earn less than whites. Men's earnings both rise with age, but at a diminishing rate, even turning down at older ages between 1980 and 2000. For women, earnings also rise with age, but not as quickly. Conspicuously, earnings rise with years of school for both men and women.

Of these patterns, a number of outcomes are particularly surprising. First, the gender gap (in percent terms) for whites is almost twice as great as that of blacks. In 1980, white women earned 58% as much as white males, but black women earned 76% as much as black men yielding gender gaps of 42 and 24%, respectively. In 2016, these figures were 75% for whites and 86% for blacks, yielding gender gaps 25 and 14%. Clearly, during this 36-year period, women's earnings rose relative to men's, such that the gender gap diminished equally by about 40% for both whites and blacks. Second, as also seen in Fig. 1, the gender wage gap starts out relative small for younger workers (24% for 20- to 24-year-olds in 1980 and only 13% in 2012), but more than doubles by the time employees reach the 54–65 age bracket. Thus, older women fair far worse relative to men than younger women. Third, the level of education has no effect on the gender wage gap. In 1980, women high school graduates and below earned about 60% of male earnings (a 40% gap) which was similar to college-educated women. In 2016, the pay ratio was about 70%, which again was similar at each level of education. So on average, women don't fare any worse with little education compared to women with college degrees. Fourth, the black–white earnings gap for men remains relatively constant, being 27% in 1980 and 25% in 2016.

Table 2 gives US results based on age and marital status. Again, earnings rise with age at a diminishing rate. However, here, the gender wage gap is far smaller for singles than marrieds. As before, the gender wage gap rises with age for marrieds, but not so much for singles. When accounting for children, the results are more stark. Table 3 indicates only a 5% gender gap in 2012 for single childless women, but a 28% gap for married men and women with children 6–17 years of age. Finally, spacing children more widely exacerbates the gender gap further (Polachek 1975b).

Taken together, we find earnings rise with education and age, differ by gender but more so for whites than blacks, and that being married and having children widely spaced apart intensifies gender earnings differences. In short, earnings disparities abound throughout the USA.

Patterns observed in the USA also hold true in other countries. The Luxembourg Income Study (LIS) contains harmonized survey microdata from over 47 upper- and middle-income countries. Tabulations in Table 4

Table 1 Average weekly earnings by gender, age, and educational groups, USA (2000 USD)

	1980		1990		2000		2010		2016	
	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
White	847	494	841	559	917	628	973	702	973	726
Black	617	467	603	510	674	555	716	586	725	625
Age										
16–24	502	383	426	377	410	374	443	372	435	377
25–34	771	514	699	547	745	594	760	621	768	644
35–44	944	520	932	610	953	652	1017	737	1040	778
45–54	953	522	994	597	1077	683	1093	752	1083	792
55–64	912	506	961	545	1075	645	1109	759	1112	773
Education										
<=8	602	353	488	336	443	334	425	319	460	313
1–3 years HS	661	391	547	367	489	357	492	345	511	347
4 years HS	763	453	688	466	670	462	654	489	669	479
1–3 years Col	841	508	814	553	829	582	823	591	782	592
4 years Col	1095	668	1152	780	1338	891	1419	963	1396	972

Note The numbers represent the average weekly earnings in 2000 USD
Source IPUMS-CPS (March rounds)

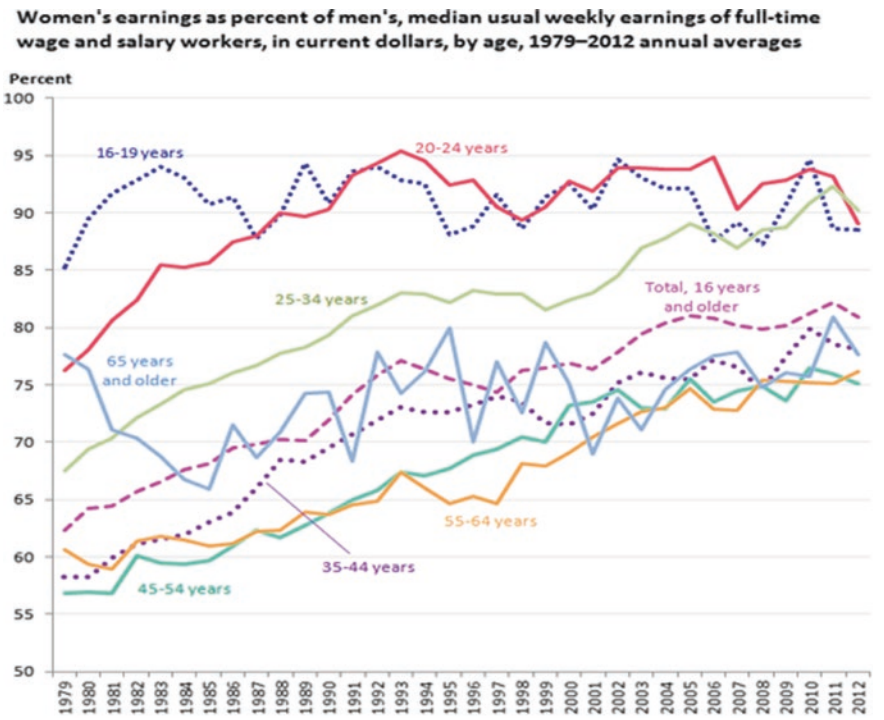

Fig. 1 Gender wage differentials (by age group) (*Source* US Bureau of Labor Statistics)

contain earnings data (2011 PPP US dollars) by age, education, and gender for 23 countries contained in the LIS. As in the USA, earnings rise with age and schooling level. Men's age-earnings profiles are steeper than women's. The gender earnings gap is smaller for the young (indeed women have the advantage in at least four countries), but rises as employees get older. Also, as in the USA, the gender earnings gap appears independent of one's years of school. Finally, Table 5 examines LIS data by gender and marital status (again in 2011 PPP US dollars). For most countries, wage parity is observed for unmarried men and women. (Exceptions are France, Israel, Japan, and surprisingly Norway and Sweden with the biggest gaps for the unmarried.) As in the USA, the gap varies from 13 to 50% for married men and women, with the largest gaps being in France, Norway, and Sweden.

In summary, earnings are not uniform across demographic groups. Instead, they differ by race, gender, age, education. Some patterns are expected, such as how earnings rise with schooling, but other patterns are not, such as how the gender earnings gap rises with age, but not years of

Table 2 Average weekly earnings by gender and marital status, USA (2000 USD)

	1980		1990		2000		2010		2016	
	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
Married, spouse present										
16–24	572	388	475	381	456	391	552	436	509	409
25–34	809	504	749	540	799	611	850	656	878	702
35–44	974	506	985	605	1040	664	1117	790	1138	853
45–54	987	502	1030	586	1139	696	1182	772	1177	830
55–64	933	492	989	541	1126	642	1176	770	1193	812
Never married/single										
16–24	459	377	407	375	399	372	423	362	422	372
25–34	658	547	628	583	687	600	674	612	688	608
35–44	693	631	746	682	732	670	775	698	818	669
45–54	696	635	798	684	777	712	841	741	813	721
55–64	711	618	745	595	778	687	899	758	893	742

Note The numbers represent the average weekly earnings in 2000 USD
Source IPUMS-CPS (March rounds)

Table 3 Median usual weekly earnings of full-time wage and salary workers, by sex, marital status, and presence and age of own children under 18 years old, 2012 annual averages

	Men	Women	Ratio (%)
Total married, spouse present	981	751	77
With children 6–17 years, none younger	1035	746	72
Total other marital statuses[a]			
With no children under 18 years	687	654	95
With children 6–17 years, none younger	790	614	78

[a]Includes never married, divorced, separated, and widowed persons
Source https://www.bls.gov/ted/2013/ted_20131203.htm

school. Also less obvious, the gender gap is almost nonexistent between single childless men and women, but large between married men and women with children.

3 Why Do Earnings Differ?

The predominant explanation entails human capital theory (Bowles et al. 2001). This theory postulates earnings power results when individuals produce human capital through inputs as parental time, schooling, on-the-job training, and perhaps a bit of luck. Its roots go back at least to early 1691 when economists began to consider the value of wealth embodied in individuals (Kiker 1966). Sir William Petty's essay "On the Value of People" written around 1655 (Hull 1899, pp. 108–112) computed the worth of people based on deducting property rent from national income. Later economists who considered human value include Adam Smith (1723–1790), Gaspar Melchor de Jovellanos (1744–1811), Jean-Baptiste Say (1767–1832), Nassau William Senior (1790–1864), Friedrich List (1789–1846), Johann Heinrich von Thünen 1783–1850), Ernst Engel (1821–1896), Léon Walras (1834–1910), and Irving Fisher (1867–1947) who formally used the term "human capital" in 1897. These economists tended to consider aggregate labor which they applied to measuring national wealth and its changes resulting from war, migration, and disease. Not until 1935 did John Walsh, and later in 1945 did Milton Friedman and Simon Kuznets, consider specific occupations. Although human capital theory evolved over a long period of time, it did not really take off until 1958 when Jacob Mincer embedded schooling into a cogent parsimonious investment framework showing precisely how years of education translate into earnings power. Slightly later, Gary Becker and Barry Chiswick (1966), Yoram Ben-Porath (1967), and then

Table 4 Average labor earnings, by countries (in 2011 PPP US dollars)

		Age group						% peak to trough	Education			
		16–24	25–34	35–44	45–54	55–64	All		Low	Medium	High	All
Australia	Male	29,550	45,253	54,382	56,128	53,832	49,040	47	36,013	42,234	66,043	49,040
Australia	Female	27,037	40,183	40,896	40,113	39,282	38,328	33	29,686	31,749	45,984	38,328
	% difference	9	11	25	29	27	22		18	25	30	22
Belgium	Male	21,215	28,899	31,405	51,641	33,873	35,882	59	25,450	41,544	39,571	35,982
Belgium	Female	16,991	21,862	23,327	27,424	24,015	23,328	38	17,713	21,234	26,426	23,341
	% difference	20	24	26	47	29	35		30	49	33	35
Brazil	Male	7156	11,761	14,350	15,751	15,972	12,988	55	8252	12,902	37,971	13,108
Brazil	Female	6572	10,247	11,733	12,198	12,243	10,535	46	5937	8193	23,468	10,722
	% difference	8	13	18	23	23	19		28	36	38	18
Canada	Male	25387	42942	53201	56991	60429	51869	55	32578	41961	58972	51877
Canada	Female	24045	34,684	40,833	44,263	38,070	39,054	46	23,841	30,883	43,214	39,100
	% difference	5	19	23	22	37	25		27	26	27	25
China	Male	8101	12,065	13,366	14,123	14,098	13,148	43	10,697	12,265	16,682	13,151
China	Female	7500	10,042	10,981	11,855	8636	10,661	37	7718	10,745	13,960	10,660
	% difference	7	17	18	16	39	19		28	12	16	19
Denmark	Male	28,689	47,169	58,254	59,995	56,576	54,882	52	41,493	49,814	68,620	54,921
Denmark	Female	26208	39,919	44,281	45,349	43,864	43,260	42	35,801	39,659	49,484	43,280
	% difference	9	15	24	24	22	21		14	20	28	21
Finland	Male	29,032	40,178	50,584	52,729	52,103	48,152	45	37263	40,046	58,976	48,152
Finland	Female	26,075	33,688	38,887	39,548	39,073	37,838	34	29,724	30,885	43,126	37,838
	% difference	10	16	23	25	25	21		20	23	27	21
France	Male	5664	22,006	28,250	29,863	16,107	13,606	81	7344	16,273	32,576	16,719
France	Female	3676	16,084	17,802	18,411	8821	8497	80	3802	10,353	21,293	10,229
	% difference	35	27	37	38	45	38		48	36	35	39
Germany	Male	30,658	44,662	56,521	59,520	61,856	56,629	48	37,558	45,866	76,708	56,719
Germany	Female	25,563	38,978	43,900	45,907	50,298	44,458	44	32,888	38,295	55,690	44,505
	% difference	17	13	22	23	19	21		12	17	27	22
India	Male	3286	4636	5013	6471	6864	5080	49	3657	6257	10,474	5081
India	Female	3315	3990	3378	3983	4180	3642	17	1982	6341	9813	3643
	% difference	−1	14	33	38	39	28		46	−1	6	28

		Age group						% peak to trough	Education			
		16–24	25–34	35–44	45–54	55–64	All		Low	Medium	High	All
Israel	Male	9123	25,763	37,273	39,320	40,999	31,232	77	18132	22,065	44,267	31,281
Israel	Female	7963	20,402	27,759	28,107	33,570	23,807	72	14,711	15,329	31,580	23,830
	% difference	13	21	26	29	18	24		19	31	29	24
Italy	Male	16,701	21,069	26,772	29,335	30,415	26,655	43	22,266	26,603	38,923	26,655
Italy	Female	14,245	18,880	24,203	24,094	27,324	23,081	41	17,692	22,007	30,624	23,081
	% difference	15	10	10	18	10	13		21	17	21	13
Japan	Male	28,236	30,808	42,148	50,411	46,794	41,707	44	25,980	36,548	48,491	41,739
Japan	Female	15171	21,728	24,315	24,831	20,923	22,019	39	13,560	18,084	27,361	21,986
	% difference	46	29	42	51	55	47		48	51	44	47
Luxembourg	Male	32,004	47,287	63,538	66,530	76,510	60,581	52	38,933	52,998	90,885	60,788
Luxembourg	Female	34,784	51,239	54,456	59,217	56,088	53,645	41	32,827	45,776	72,875	53,859
	% difference	−9	−8	14	11	27	11		16	14	20	11
Mexico	Male	4861	7944	9966	9988	9276	8184	51	5402	9136	21,768	8184
Mexico	Female	4086	7115	8408	7751	7812	6876	47	4112	7172	14,984	6876
	% difference	16	10	16	22	16	16		24	21	31	16
Netherland	Male	28,009	43,847	61,002	64,935	68,314	58,559	57	41,437	48,616	76,043	58,754
Netherland	Female	28,273	43,464	58,108	55,782	50,329	49,940	49	32,334	42,621	57,899	50,332
	% difference	−1	1	5	14	26	15		22	12	24	14
Norway	Male	12,111	37,851	51,396	54,152	45,658	28,186	78	17,731	35,584	54,512	35,836
Norway	Female	8717	24,294	33,225	35,275	26,673	17,199	75	9243	19,273	34,767	21,724
	% difference	28	36	35	35	42	39		48	46	36	39
Poland	Male	2210	11,746	13,902	10,080	5934	6035	78	1401	7324	16279	7356
Poland	Female	1342	6778	8287	7061	2919	3525	81	411	3441	10,444	4142
	% difference	39	42	40	30	51	42		71	53	36	44
Russia	Male	12,287	16,410	17,282	14,860	12,337	15,169	17	12,208	13,335	18,064	15,171
Russia	Female	9400	10,928	11,783	11,520	9801	10,972	18	8440	9338	11,988	10,972
	% difference	23	33	32	22	21	28		31	30	34	28

(continued)

Table 4 (continued)

		Age group						% peak to trough	Education			
		16–24	25–34	35–44	45–54	55–64	All		Low	Medium	High	All
Spain	Male	15,498	24,626	32,559	36,750	38,481	33,037	58	24,104	30,456	42,500	33,038
Spain	Female	12,758	23,489	27,429	30,984	31,532	28,145	59	17,915	22,444	34,480	28,153
	% difference	18	5	16	16	18	15		26	26	19	15
Sweden	Male	7951	26,409	35,275	35,130	28,926	18,109	77	13,088	26,048	36,915	25,016
Sweden	Female	5941	16,624	22,262	24,818	19,359	11,632	76	7102	16,329	24,228	16,461
	% difference	25	37	37	29	33	36		46	37	34	34
UK	Male	22,111	36,946	45,892	47,164	42,491	41,008	53	26,629	33,709	53,678	41,119
UK	Female	22,189	32,079	39,989	37,357	34,305	34,184	41	22,191	26,572	41,203	34,206
	% difference	0	13	13	21	19	17		17	21	23	17
USA	Male	30,904	50,043	68,448	73,839	76,422	65,118	58	34671	47,260	88,022	65,118
USA	Female	25807	42,998	49,132	50,593	51,197	47,156	49	25,643	35,742	57,719	47,156
	% difference	16	14	28	31	33	28		26	24	34	28

Note The calculations are based on LIS person-level survey data for different countries in different years. All averages are at constant prices (2011) and expressed in 2011 PPP US dollars. For Australia, Canada, Denmark, France, Italy, the averages are for 2010; for Belgium the averages are for 2000; for Brazil, Finland, Germany, Luxembourg, Netherland, Norway, Poland, Russia, Spain, UK, USA, the averages are for 2013; for Israel, Mexico, the averages are for 2012; for Sweden the averages are for 2005; for China the averages are for 2002; for India averages are for 2011; for Japan averages are for 2008. For Australia, Brazil, China, Israel, Japan, Mexico, Russia, UK, the calculations are based on workers who worked at least 35 hours per week; for Belgium, Canada, Finland, Germany, Italy, Luxembourg, Netherland, Spain, USA, the calculations are based on full-time workers who worked at least 35 hours per week; for Denmark, India, the calculations are based on workers who identify themselves as full-time workers; for France, Norway, Poland, Sweden, the calculations are based on all workers as no information on work intensity is available

Education categories Low: no education, pre-primary, primary, lower secondary education, compulsory education, initial vocational education; Medium: upper secondary general education, basic vocational education, secondary vocational education, post-secondary education; High: specialized vocational education, university/college education, (post)-doctorate and equivalent degrees

Source LIS datasets

Table 5 Average labor earnings, by countries (in 2011 PPP US dollars)

Country		Unmarried[a]	Married
Australia	Male	39,150	53,317
Australia	Female	36,584	39,391
		0.93	0.74
Belgium	Male	26,006	39,847
Belgium	Female	22,551	23,008
		0.87	0.58
Brazil	Male	9538	16,314
Brazil	Female	8866	12,281
		0.93	0.75
Canada	Male	37,965	55,504
Canada	Female	37,214	39,481
		0.98	0.71
China	Male	9513	13,563
China	Female	8646	10,900
		0.91	0.80
Denmark	Male	44,992	60,304
Denmark	Female	40,626	44,111
		0.90	0.73
Finland	Male	40,098	53,305
Finland	Female	35,188	39,316
		0.88	0.74
France	Male	8893	19,815
France	Female	7059	11,229
		0.79	0.57
Germany	Male	47,753	60,477
Germany	Female	42,542	46,145
		0.89	0.76
India	Male	4271	5293
India	Female	5000	3391
		1.17	0.64
Israel	Male	17,452	36,809
Israel	Female	14,735	27,239
		0.84	0.74
Italy	Male	21,763	28,483
Italy	Female	22,022	23,496
		1.01	0.82
Japan	Male	30,352	43,727
Japan	Female	22,229	21,940
		0.73	0.50
Luxembourg	Male	50,986	62,921
Luxembourg	Female	52,562	54,068
		1.03	0.86
Mexico	Male	6416	8701
Mexico	Female	6655	6980
		1.04	0.80

(continued)

Table 5 (continued)

Country		Unmarried[a]	Married
The Netherlands	Male	48,355	63,600
The Netherlands	Female	46,096	55,547
		0.95	0.87
Norway	Male	18,481	42,371
Norway	Female	12,358	24,595
		0.67	0.58
Poland	Male	4131	9640
Poland	Female	3451	5287
		0.84	0.55
Russia	Male	13,265	16,161
Russia	Female	11,994	10,284
		0.90	0.64
Spain	Male	25,974	34,959
Spain	Female	24,950	29,035
		0.96	0.83
Sweden	Male	22,016	26,713
Sweden	Female	15,793	16,562
		0.72	0.62
UK	Male	30,974	43,954
UK	Female	31,371	35,331
		1.01	0.80
USA	Male	45,385	73,602
USA	Female	41,191	50,473
		0.91	0.69

[a]Never married
Source LIS datasets

Jacob Mincer (1974) extended the human capital model to incorporate work experience obtained over the life cycle. Even later, Finis Welch (1974), Chiswick (1978), and George Borjas (1982, 1985, 1993) applied the model to analyze race, ethnicity, and country of origin. Finally, Solomon Polachek (1975a) modified the model to understand gender differences. Of course, many other factors besides human capital can influence individual earnings. These comprise institutional factors including unions, market structure, government legislation, discrimination, corporate payment schemes to enhance productivity, as well as individual factors such as non-cognitive personality traits.

Before the human capital approach became popular, the predominant theory of earnings distribution attributed success mostly to luck. Earnings were depicted by purely stochastic nonrandom walk models. One approach (Roy 1950) argues that individuals possess various characteristics, each is independent of each other and distributed approximately normally across the population. By the central limit theorem, output and hence income will be log-normal if production can be depicted as the product of each

characteristic and there are a large number of characteristics. Such a theory offers no economic rationale into the earnings generation process because the individual characteristics have no behavioral content (von Weizsäder 1993); in short, earnings are the product of random variables.

Since the development of human capital theory, other models evolved to consider various factors that affect earnings. These include occupational segregation models, crowding models, efficiency wage models, matching models, and models depicting productivity-enhancing contracts. Occupational segregation and crowding models describe why women's and minority groups' outcomes differ from the majority group's, based on differences in the occupational distribution of men and women, blacks and whites, as well as other demographic groups (Bergmann 1971, 1974). Presumably, discrimination motivates employers to pigeonhole job applicants into selected occupations based on preconceived notions, often referred to as statistical discrimination, namely the stereotyping of individuals based on aggregate group perceptions. Such categorization can lead to crowding of women and minorities into particularly low-paying more menial occupations. The increased supply of workers in these occupations exacerbates a downward pressure on wages, and the resulting shortage of employees in the so-called good occupations raises wages there. Clearly, the power of this theory depends upon how occupation is defined. Defining occupations broadly weakens its explanatory power, whereas defining occupations narrowly strengthens it (Polachek 1987). Further, the theory does not explain why the gender wage gap widens with age or why family characteristics such as marital status and children are related to earnings in the opposite way for men and women, or why the gender gap is smaller for blacks than whites.

Efficiency wage models argue that some individuals earn more than competitive market wages. Paying more than the competitive wage is said to enhance productivity. First, employees work harder and shirk less for the fear of being laid off because, at best, a layoff results in a rehire at or below the competitive wage, especially if a layoff signals weak job performance. Second, employers better screen heterogeneous job applicants, choosing to employ only the best. Whereas such models justify unemployment, they don't explain why efficiency wages vary over the life cycle or why earnings vary by race and gender.

Matching models explain how wages rise over the life cycle as worker–firm combinations improve in quality and as turnover declines with tenure (Jovanovic 1979; Hosios 1990). However, they are weak in explaining the rate at which earnings rise and the periodicity of labor turnover (Polachek 2012; Polachek and Horvath 2012). Further, they don't explain gender, racial, or other demographic earnings differences.

Productivity-enhancing contract models design compensation packages to maximize employee performance not just at a point in time, but throughout a worker's potential life with a given employer. By providing appropriate incentives, such pay schemes have implications regarding earnings over a worker's life cycle. For example, piece-rate pay and bonuses increase employee pay and hence effort within a given timeframe. Promotions, often modeled as a rank-order tournament (Lazear and Rosen 1981), increase future pay as a reward for earlier efforts and as a result imply upward sloping earnings profiles (Lazear 1995). However, they give little insight regarding earnings profile concavity, and indeed, some models predict convex earnings over the life cycle as might be the case for superstars and CEOs (Rosen 1981). To get at gender wage differences, such models rely on statistical discrimination in that firms predict women on average work less over their lives, thus requiring a higher ability than for men to achieve comparable promotions to enable firms to recoup specific training costs (Lazear and Rosen 1990). However, the problem with this approach is it predicts relatively more able women than men move up the job ladder which should imply higher female productivity and hence greater female wages at more senior jobs, an observation that does not appear to be the case in the data.

Each of the above theories offers some insight for particular aspects of earnings. However, none give a unified framework that explains each of the observed earnings patterns illustrated above. We believe only the life-cycle human capital model accounts for the preponderance of all patterns simultaneously. Thus, we focus on human capital theory and the empirical work emanating from it.

4 The Human Capital Model

The backbone behind formal structural human capital models originates with Adam Smith (1776). He argued job characteristics shape labor market equilibria because workers need to be compensated for taking "unpleasant" jobs. Though going to school and investing in on-the-job training need not be unpleasant, these activities typically take time away from paid work and for this reason yield a wage premium.[3] As such, the extra money needed to forgo pay while undertaking human capital purchases is a "compensating wage

[3]Human capital investment comprises of general and specific training. Specific training enhances productivity in the firm and nowhere else. Firms provide such training because of its limited applicability. However, to reduce turnover, incentive compatible contracts can arise in which firms equally share with its employees the costs and benefits of such training (Kuratani 1973). This survey deals mainly with

differential." Couched in an investment framework, this means the present value of earnings an individual needs to obtain must exceed the costs of such expenditures, of course including direct and indirect opportunity outlays.

At first, only schooling investments were considered (Mincer 1958 and Becker 1964), but rigorous lifetime models (Ben-Porath 1967) imply something more. Assuming a finite working life and opportunities for post-school investments, such as through on-the-job training, individuals have an incentive to invest throughout their lives, but at a diminishing rate. Large human capital investments during school, followed by gradually diminishing human capital investments throughout the life cycle, lead to the typically observed concave earnings profile. Ben-Porath derived this result by assuming individuals invest in themselves to maximize lifetime earnings subject to their initial human capital (E_0) and the production technology associated with further human capital creation:

$$\text{Max}_{K_t} J = \int_0^N Y_t e^{-rt} \mathrm{d}t \tag{1}$$

where J is the total discounted disposable earnings over the working life cycle, r is the personal time discount rate, and N is the number of years after which one retires (assumed known with certainty).[4] Disposable earnings are $Y_t = R[E_t - K_t]$ where R is the rental rate for human capital E_t, and K_t is the fraction of human capital stock reinvested. Individuals create human capital using various inputs. This activity can be modeled using a very general production function, but for simplicity most employ a Cobb-Douglas model combining own time and existing human capital. Typical studies denote the human capital accumulation process as $Q_t = \beta K_t^b$ where b and β are production function parameters.[5] The parameter b indicates the rate

general training which enhances productivity throughout the economy. The cost of general training is usually borne by the individual, though because of its social value, much of schooling is subsidized by the government. Bishop (1997) presents evidence that employers can pay for general (as opposed to only specific) training. Acemoglu and Pischke (1999) argue firms can pay for general training because even general training can have a specific component. In this chapter, we concentrate on the costs and benefits of that part of human capital an individual purchases either in school or on the job.

[4]Initial human capital is determined by genetics as well as initial parental and other investments. We examine parental investments later in this chapter.

[5]Ben-Porath (1967) assumed a more general production function employing "goods" inputs such as teachers and books $q_t = \beta K_t^{b_1} D_t^{b_2}$ where D_t equals other inputs. Because goods inputs are difficult to measure, most analyses subsequent to Ben-Porath omit this factor. These include Haley (1976), Johnson (1978), and Wallace and Ihnen (1975).

at which current human capital stock is transformed to new human capital. It reflects how one acquires new knowledge from old and as such exhibits how quickly one learns. The β parameter depicts the "scale" at which one learns and hence represents total factor productivity. Both β and b reflect an individual's ability to learn. An individual's initial human capital stock (E_0) becomes relevant when determining K_t during the process of lifetime earnings maximization. One can interpret E_0 to be a person's initial ability to earn. The rate of change in human capital stock E_t is expressed as the amount of human capital produced Q_t minus depreciation so that $\dot{E} = Q_t - \delta E_t$ where δ is the constant rate of human capital stock depreciation. This depreciation parameter is symbolic of one's ability to retain (or not retain) knowledge.[6]

Maximization of lifetime earnings requires producing human capital to equate its marginal costs and marginal benefits in each time period. This yields a complex nonlinear (in the parameters) earnings function that expresses earnings in terms of school and labor market experience, and contains parameters ($E_0, \beta,$ and b) as well as discount (r) and depreciation rates (δ) as identifiable coefficients.[7] Given its complexity, only a handful of studies estimate this (or related) earnings function. These include Johnson and Hebein (1974), Rosen (1976), Haley (1976), Heckman (1976), Heckman et al. (1998), Theeuwes et al. (1985), Song and Jones (2006), Wu (2007), and Liu (2009), all of whom do so by aggregating across the entire population. An alternative theoretical version designed to get at some aspects of heterogeneity is given in Magnac et al. (2018).

Three issues underlie this earnings equation. First, the derivation assumes a continuous employment history. However, not all individuals work continuously throughout their career. This is particularly the case with women. According to Polachek (1975a) and later Weis and Gronau (1981), discontinuous (also referred to as intermittent) work implies a non-monotonic decline in human capital investment over one's lifetime. To date, no one to our knowledge has derived the resulting earnings function for such

[6]Parameters b, β, r, δ, and E_0 are assumed constant throughout one's life. Obviously, this need not be the case, but is consistent with the notion that IQ remains constant (Tucker-Drob 2009). Of the parameters, skill depreciation seems most likely to increase as one ages, but to our knowledge, no one has estimated how skill depreciation increases with age in the context of a life-cycle human capital model.

[7]A derivation of the exact function is given as Eq. (7) in Polachek et al. (2015) and derived in their Appendix 1. Their specification differs slightly from Haley (1976) in that it assumes a two-term Taylor expansion for the third term in Haley's earnings function, thus enabling them to identify all five earnings function parameters. We present this equation in Appendix 1 because it is used later in this chapter to assess the importance of ability E_0, β, b compared to schooling and experience which is used in more traditional approaches to explain earnings.

a discontinuous worker. Second, at the time it was initially derived, the human capital life cycle earnings function model proved difficult to estimate given its complex nonlinear structure. As such, most analyses adopted a quadratic approximation derived by Mincer (1974). Third, because its derivation is based on an individual's optimization of lifetime earnings, one should estimate the earnings function using lifetime data for a given person. Instead, at least until recently, all analyses used cross-sectional or panel data to obtain aggregate population-wide estimates.

Polachek et al. (2015) and Verdon (2018) exploit panel data to estimate separate equations by individual. Doing so enables one not only to get at heterogeneity, but also to test, at the time, previously unverified theorems of the life-cycle human capital model, which we will explain further in Sect. 6.

5 Simplification of the Earnings Function

Estimation of nonlinear functions derived from a life-cycle model is difficult because the equation's complex nonlinear specification impedes convergence. Polachek et al. (2015), hereafter PDT, utilize the Genetic Algorithm to optimize numeric strings using genetic reproduction, crossover, and mutation concepts (Goldberg 1989).[8] These techniques globally search the parameter space leading to convergence more efficiently than traditional Newton-Raphson hill-climbing algorithms which rely on a point-to-point gradient-based search (Dorsey and Mayer 1995). Even so, the technique is computationally time-intensive, especially when estimating the earnings function person by person.

The complexity of estimating these nonlinear equations is probably why most analyses use a simplified formulation based on Mincer (1958, 1974). Further, because long enough panel data were not available, all prior analyses estimated aggregate population-wide earnings functions. Given this extensive research, we now examine Mincer's formulation as well as various extensions of it. This entails describing his specification and interpreting its implications. Following this, we use Mincer's results as a benchmark to evaluate what can be learned from obtaining individual-specific parameters. Then finally, we deal with techniques current researchers use to get at exogeneity issues regarding returns to human capital investment.

[8]The algorithm was originally developed by Holland (1975). PDT use a version of GA written by Czarnitzki and Doherr (2009).

5.1 The Mincer Model

By equating the present value of school investment benefits to its costs, Mincer (1958) derived his original earnings function

$$\ln Y_i = \alpha_0 + r_s S_i \tag{2}$$

where Y is earnings and S is schooling, which he estimated using cross-sectional US census data. Of prime interest was the r_s coefficient that depicts the rate of return to school. Of less interest is α_0 that represents the logarithm of earnings assuming no school. Even in the 1960s when Mincer first estimated this equation, he realized the equation had shortcomings. Most obvious was an omitted experience variable which is necessary in order to introduce life-cycle considerations into the model. This omission causes r_s to underestimate the true rate of return because both schooling and labor market experience are positively related to earnings, but schooling and labor market experience are inversely correlated.[9] Becker and Chiswick (1966) as well as Mincer (1974)[10] incorporate Ben-Porath's (1967) theorem that human capital investments decline monotonically with age assuming a finite (and continuous) work horizon. This inclusion yields a concave earnings function. Although Mincer experimented with several specifications,[11] the following log-linear model is the one that prevailed, probably because of its simplicity

$$\ln Y_i = \alpha_0 + r_s S_i + \beta_1 t_i + \beta_2 t_i^2 + \epsilon_i \tag{3}$$

where all variables are the same as before, except now t represents work experience.[12] The coefficient α_0 is related to initial earnings capacity, and β_1 and β_2 are a combination of the amount and the return to human capital

[9]Of course, there were other biases but these were considered later.

[10]Also Tom Johnson (1970).

[11]These include a Gompertz specification as well as various interaction terms.

[12]Murphy and Welch (1990) experiment with cubic and quadratic functional forms. Heckman and Polachek (1974) use Box-Cox and Box-Tidwell transformations to show the log-linear fit works best when compared to a set of other common functional forms. Heckman et al. (2003) modify the Mincer model to incorporate individuals choosing their education levels to maximize their present value of lifetime earnings. They also relax other restrictions such as the constraint that log earnings increase linearly with schooling and the constraint that log earnings-experience profiles are parallel across schooling classes, but Mincer also relaxes these latter constraints in a number of his specifications which contain an interaction term between experience and schooling. Indeed, he finds (1974, pp. 92–93) nonparallel profile shifts, as well.

investments.[13] Numerous examples of this equation appear in the literature. All yield positive returns to schooling (in the 3–20% range) and all yield concave earnings profiles (exhibited by negative β_2 coefficients), but here too, there are biases.

Mincer estimated (3) using the 1960 Public Use US Census data to obtain:

$$\ln Y = 6.20 + .107\,S + .081\,t - .0012\,t^2 \tag{4}$$

Given there are four coefficients representing five aspects of human capital, one must make an identifying restriction.[14] Assuming equal rates of return for schooling and post-school investment ($r_s = r_t$) yields an E_0 of \$1185.59 in 1960 dollars, or \$9778 in 2016 dollars, which reflects the earnings power of an individual with no human capital. The initial time-equivalent investment when just completing school (k_0) equals 0.492 meaning that one initially spends about 50% of the time on one's first job investing in on-the-job training. Finally, T equals 25.82, implying that earnings peak just after 25 years in the labor force.[15]

According to the Ben-Porath optimization model, human capital investment declines continuously over one's lifetime. If going to school entails 100% use of one's time, then time investment just after completing school should be slightly below 1.0, but not as low as the 0.5 observed above. One explanation centers on governmental and familial subsidies to those attending school (Johnson 1978). According to this argument, school enrollees receive subsidies if and only if they remain in school. To obtain the maximum subsidy, individuals stay in school longer than otherwise (a distortion), but revert back to the non-subvention investment patterns when the subsidy

[13]These five aspects are related to, but not exactly the same as, PDT's five parameters. The coefficients $\alpha_0 = \ln E_0 - k_0[1 + \frac{k_0}{2}]$, $\beta_1 = r_t k_0 + \frac{k_0}{T}(1 + k_0)$ and $\beta_2 = -\left[\frac{r_t k_0}{2T} + \frac{(k_0)^2}{2T}\right]$ assuming a linearly declining post-school investment function $k_t = k_0 - \frac{k_0}{T}t$ where k_0 is initial and k_t concurrent "time-equivalent" investment and T is the total period of positive investments. Mincer also considered three other specifications for k_t. These entail (1) a linear declining dollar specification, (2) an exponentially declining dollar specification, and (3) an exponentially declining time-equivalent investment specification. These yielded nonlinear in the parameters less popular earnings functions that by and large have been ignored in the literature.

[14]The parameters are the initial human capital stock (E_0), the rate of return to schooling (r_s), the rate of return to post-school human capital investment (r_t), and the time when gross human capital investment just equals depreciation which is the experience level at which net human capital investment goes to zero (T).

[15]The computation results from solving the following equations:

$$\ln E_0 - k\left(1 + \tfrac{k}{2}\right) = 6.2;\ r_s = .107;\ r_t k + \tfrac{k}{T}(1 + k) = .081;\ -r_t \tfrac{k}{2T} + \tfrac{k^2}{2T^2} = -.0012;\ r_s = r_t;\ \text{for } T, k, r_s, r_t \text{ and } Y.$$

disappears. Given possible social benefits from an educated population, this seemingly overinvestment in school is not necessarily suboptimal from a national perspective.[16]

5.2 Direct Applications of the Mincer Earnings Function

At least three important empirical implications emerge directly from the Mincer earnings function. First, earnings rise with human capital investments. This means the coefficient on schooling should be positive, and it is bigger the better the quality of education. Second, the coefficient on experience-squared should be negative indicating less earnings growth mid-career. Third, earnings distribution should be related to both levels and variations in human capital accumulation. This means the variance of earnings widens as schooling levels increase and as a population ages. However, interestingly, holding schooling level constant, relative earnings differences (as measured by the variance of the logarithm of earnings) should narrow with experience and then widen, exhibiting a U-shaped log variance of earnings (Polachek 2003).

Each of these is widely observed. Literally, dozens of studies estimate schooling rates of return. These entail multiple countries and cover numerous years. One survey (Patrinos and Psacharopoulos 2010) contains rate of return estimates for over 70 countries spanning more than 25 years. A second (Trostel et al. 2002) contains estimates for 28 countries. A third (Montenegro and Patrinos 2014) utilizes the World Bank International Income Distribution Database to estimate rates of return for 139 economies mostly since 2000. In a meta-analysis using 97 rate of return estimates from 27 studies, Orley Ashenfelter et al. (1999) find "little controversy … schooling adds considerably to the earnings of individuals," that "rates of return to schooling appear to be higher in the USA than elsewhere (p. 466)," and these returns have increased between 1980 and 1999. Philip Oreopoulos and Uros Petronijevic (2013) in a survey on the returns to college education also claim that "the earnings premium associated with a college education has risen substantially" and that college is still a "sound investment" (p. 1).

Although more school is associated with higher earnings, it is not obvious schooling actually raises productivity. A number of theories claim not. For example, signaling models argue that better workers "signal" their prowess by

[16]See Psacharopoulos and Patrinos (2004) and Psacharopoulos (2006) for an analysis of social rates of return to education.

going to school, but school itself doesn't affect productivity. Similarly screening models claim that firms screen on certain characteristics such as completing a degree because "finishing" signals stick-to-itiveness a characteristic defining potentially "better" workers, but again schooling by itself doesn't affect productivity. Finally, long-term contract models yield escalating life-cycle earnings. However, these pay schemes reflect techniques firms use to hire the best workers, decrease turnover and minimize job shirking, but do not necessarily increase worker productivity. Although actual employee productivity is hard to measure, and few data sets actually have such information, some studies exist linking educational investments to actual productivity. For example, utilizing productivity data on 296 household farms in West Bengal, India, Kumbhakar (1996, p. 188) showed "that education increases [actual] productivity" and that this enhanced productivity increased farmers' wages. Generalizing these results to economic growth, Barro and Sala-i-Martin (1999) find that the higher the population's education, the higher its GDP and GDP growth per capita. Related to actual productivity, Craig Riddell and Xueda Song (2017) using Canadian data find education increases the probability a worker will be using a computer on the job. With regard to sheepskin effects, Clark and Martorell (2014) find little evidence of signaling when comparing the earnings of workers who barely passed and barely failed exams leading to a high school diploma. With respect to social effects of school, Lochner and Moretti (2004) show that schooling reduces the probability of incarceration and arrest. In another realm, Benmelech and Berrebi (2006), based on a unique data detailing the biographies of Palestinian suicide bombers, find that more educated suicide bombers are more likely to succeed in their mission and are more likely to induce casualties when they attack. In addition, education positively affects non-labor market activities. For example, Michael (1973) shows that education improves one's efficiency in consuming everyday commodities. Polachek and Polachek (1989) illustrate "reverse intergenerational transfers" by showing that even one's children's education positively affects the way one consumes. In summary, schools appear to increase cognitive and non-cognitive skills. However, not obvious is whether these acquisitions primarily come about because of school or simply because of students' innate abilities. More on this later.

Also universal is earnings function concavity exhibited by a negative β_2 coefficient found when estimating Eq. (3). Early studies (Mincer 1974) tested this proposition using OLS regression with cross-sectional data.[17] This

[17]Some use panel data, but one can question how these adjust for price changes. Another exception is in executive pay late in some individuals' career paths.

result is universal across countries and years (Polachek 2008). These results also hold when one adjusts for selectivity biases (Hartog et al. 1989; Kiker and Mendes de Oliveira 1992; Baldwin et al. 1994; Gibson and Fatai 2006) and for individual-specific heterogeneity using standard and not so standard fixed-effects techniques (Mincer and Polachek 1978; Licht and Steiner 1991; Kim and Polachek 1994; Light and Ureta 1995; Bhuller et al. 2014).

Finally, as Mincer predicts, the distribution of earnings varies over the life cycle. According to Mincer, $\sigma^2(\ln Y_{it})$ where i denotes an individual and t denotes an experience level is likely U-shaped over t, with the trough occurring near Mincer's "overtaking" point" $1/r_s$ years after finishing school.[18] Predicting this trough is unique to the human capital model. At the overtaking point $(1/r_s)$, observed earnings (Y_S) are equivalent to potential earnings (E_S) from school alone, because at that point in life, the cost of and the returns to on-the-job training cancel each other out, and this is true for all individuals independent of the amount of training. Here, all earnings variation is simply $\sigma^2(E_S)$. However, earlier in one's career, say when one just leaves school, earnings variation arises from both earnings variation in potential earnings $\sigma^2(E_S)$ and variation in the cost of training. Later in life, as annual training declines, earnings variation amounts to the variations in both original potential earnings $\sigma^2(E_S)$ and the returns to past on-the-job post-school training. Mincer verified this U-shape with US data, Brown (1980) also found some evidence of this, and Polachek (2003) corroborated this with LIS data based on nine countries.

5.3 Extending the Mincer Earnings Function

Adding categorical dummy variables to the basic Mincer earnings function yields estimates of earnings differences across population subgroups. In this vein, numerous studies proliferated beginning with analyses of the union/nonunion wage gap (Lewis 1963, 1986), race (Welch 1974), gender (Fuchs 1967; Suter and Miller 1973) migration and ethnicity (Chiswick 1978; Borjas 1982, 1985, 1993), and health status (Grossman 1972). Nowadays, a host of other factors related to earnings are considered. For example, beauty (Hamermesh and Biddle 1994; Scholz and Sicinski 2015), height (Lundborg et al. 2014), dress (Hamermesh et al. 2002), hair color (Dechter 2015), grooming (Robins et al. 2011), sexual orientation (Sabia 2015; Klawitter 2015), college major (Webber 2014), bilingualism (Saiz and

[18]Proof is given in Mincer (1974, p. 103).

Zoido 2005), social skills (Weinberger 2014), personality (Groves 2005), mental state (Cseh 2008), childhood disorders (Fletcher 2014), teenage drug use (Burgess and Propper 1998), veteran status (Gabriel 2016), religion (Steen 2004), and more.

Interpreting these earnings differences is tricky as many of these variables might not be truly exogenous. This is certainly the case for schooling. If higher ability students go to school longer, then part of the often measured return to school may be a return to student ability, and not school per se. A long literature spells out and attempts to correct for this endogeneity bias arising from omitted ability. We will discuss this later. But it is also the case that other seemingly more likely exogenous variables are not truly exogenous.

Take the case of gender. Many define gender wage differences holding education, experience (though most studies use potential rather than actual experience), and other variables constant to constitute discrimination. Such regression models indicate women earn less than men. In the USA, the gap is approximately 22%. Among OECD countries, the gap averages 15%. One might argue this indicates rampant discrimination, namely that firms pay women lower wages despite seemingly equal qualifications. But the story is far more complicated.

An exogenous variable must be randomly assigned independent of other variables. Certainly, in the USA and OECD countries where there is no apparent child preference, gender is typically thought to be randomly assigned at birth. True, gender does not appear to affect or be affected by other variables in the human capital model. However, there still are a number of endogeneity issues. For one, gender is not independent of other omitted variables, but instead it is correlated with other confounding factors that may affect earnings. Expected lifetime labor force participation is the most notable. For example, marriage and motherhood are often cited as the prime reasons for intermittent participation. Women who do not get married or have children have comparable lifetime work histories and wages relative to non-married childless single men. But married women with children have wildly different lifetime labor force participation than their men counterparts.

To see the effects of these omitted variables, we modify the Mincer earnings function to include marital status and number of children, along with a set of interaction terms between these and gender. One such specification is:

$$\begin{aligned}\ln Y_i(t) = {} & a_0 + a_1 S_i + a_2 t_i + a_3 t_i^2 + \alpha_5 F_i + \alpha_6 M_i + \alpha_7 F * M_i \\ & + \alpha_8 C_i + \alpha_9 F * C_i + \alpha_{10} F * M * C_i + \alpha_{11} X_i + \varepsilon_i\end{aligned} \tag{5}$$

where $\ln(Y)$ is the logarithm of earnings, S represents years of schooling, t and t^2 depict years of experience and its square, as have already been defined;

and F is a categorical gender dummy variable for being female, M a categorical dummy variable for marital status, $F * M$ an interaction term between gender and marital status, C the number of children, $F * C$ an interaction term between gender and number of children, $F * M * C$ a three-way interaction term, X other relevant exogenous variables, and ε_i a random error term for each individual observation. This specification yields estimates of the gender wage gap for married men and women separately from single men and women. It also estimates the effect of children on the gender wage gap. The interesting result is a "family wage gap" in which the gender difference in earnings is relatively small for single men and single women, yet large for married men and married women, and especially large for those married men and women with children. Polachek (1975b) was the first to find the "family wage gap" for US data. Later, Sanders Korenman and David Neumark (1992) also documented the gap for the USA. Francine Blau and Lawrence Kahn (1992) corroborate marital status differences using international data, as does Polachek (2008) who presents results for 14 countries using the LIS data. Independent of country or year, the gender gap for singles varies between 20% in favor of men and 4% in favor of women (the unweighted average is about 8% in favor of single men over single women) to between 3 and 56% (with an unweighted average of about 30%) for married men and over married women. This means the gender wage gap is not uniform. It is small for childless single men and women, but relatively large for married men and women with children. Why?

The reason is an omitted variable. To see this note that marriage and children are related to lifetime labor force participation, but both marriage and children influence lifetime work differently for men and women. For men, being married having children is associated with higher lifetime work, but for women marriage and children decrease lifetime work. These work patterns are illustrated in both cross-sectional data and retrospective work histories. Figure 2 depicts gender–marital status labor force participation patterns for the USA in 1970 and 2010. Married men in 1970 have the highest lifetime labor force participation. Married women have the lowest, peaking at about 47% between the ages of 20 and 24. The drop between ages 25 and 35 reflects intermittent labor force participation related to childbearing. The gap between single males and females is the narrowest. By 2010, the gender differences are appreciably smaller, but still remain. Figure 3 shows how female labor force participation decreases with children. It indicates younger children have a larger negative effect on work.

The same lifetime work patterns emerge from retrospective data. Using the 1980 Panel Study of Income Dynamics Data (PSID), Miller (1993) finds that married women average 10.04 years out of the labor force.

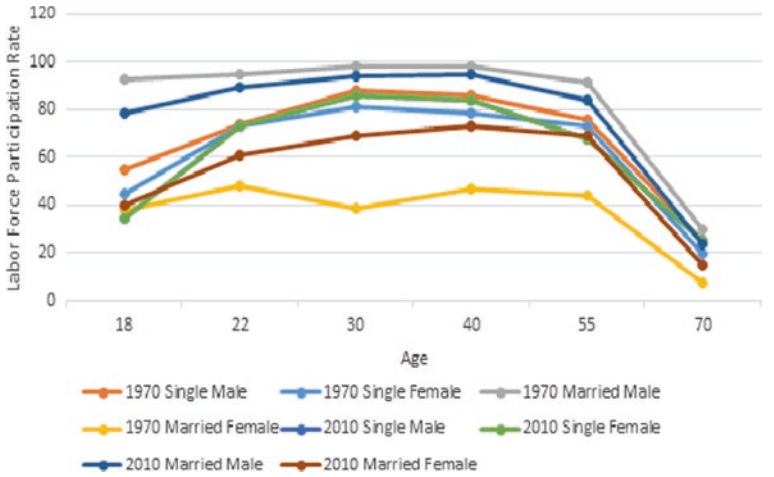

Fig. 2 US labor force participation rate (by gender, marital status, age) (*Source* US Bureau of Labor Statistics)

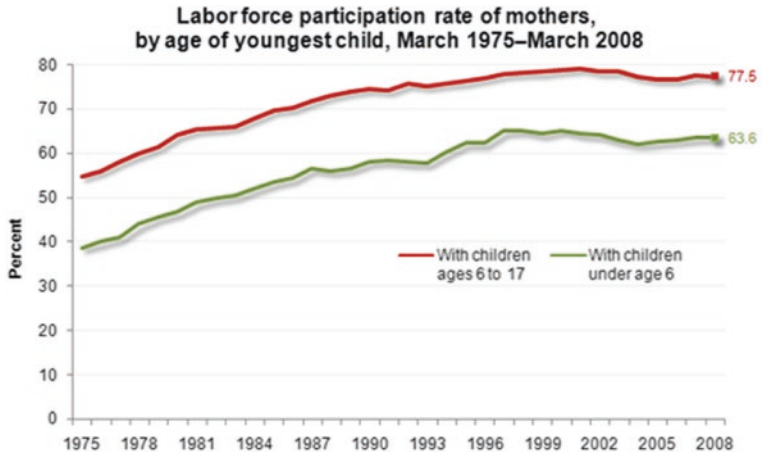

Fig. 3 Labor force participation rate of mothers (*Source* US Bureau of Labor Statistics)

Similarly, using a panel of 2659 individuals from the 1976 to 1987 PSID data, Kim and Polachek (1994) find that women averaged 9.62 years out of the labor force relative to men's 2.22 years. Current data for foreign countries are comparable. Using Canadian data, Simpson (2000) finds that in 1993 married women with children averaged 7.6 years (or 36.4% of their work years) out of the labor force, whereas single women spent 1.5 (or 12.9%) of their work years out of the labor force. For men, this figure is

0.9 years (or 8.1%). Data within narrow professions yield similar results. Catalyst (2003) finds that only 29% of women MBA graduates worked full-time continuously since graduation compared to 69% for men, and similarly only 35% of women law graduates worked continuously since graduation compared to 61% for men.

The Segmented Earnings Function

Mincer and Polachek (1974) modified earnings function (3) to incorporate discontinuous labor force participation.[19] They did so by dividing Eq. (3)'s potential experience (t) into its components: actual experience and time out of the labor force. Whereas many work and non-work segments can be used, for simplicity here we adopt two work segments (e_1 and e_2) and one "home-time" segment (h). The empirical specification is derived assuming linearly declining human capital investments in each work/non-work segment to obtain

$$\ln Y_t = a_0 + r_s S + \alpha_1 e_1 + \delta_h h + \alpha_2 e_3 + \varepsilon \tag{6}$$

where e_1, h, and e_3 are the work and non-work segments.[20]

The α_1 and α_2 coefficients range from 1.2 to 4.0%, depending on the population subgroup studied and on one's level of education. The δ coefficient ranges from −4.5 to −0.5% depending on the respondent's amount and type of education. In general, the higher one's education and the more skilled one's job, the greater the magnitude of these coefficients. Also, α_2 often exceeds α_1 because upon reentering the labor one has a greater commitment to working more continuously (Polachek 1975a). By now, numerous studies adopted this approach to assess the effect of work interruptions. Examples include Albrecht et al. (1999), Baum (2002), Corcoran and Duncan (1979), Corcoran et al. (1983), Hotchkiss and Pitts (2003, 2005),

[19]In empirical work, Mincer and Polachek (1978) adjust for endogenous lifetime work using two-stage least-squares estimation. Also see Gronau (1988).

[20]Assuming a linear human capital investment function $k(t) = a_i + b_i t$ where a_i is the initial "time-equivalent" investment and b_i is the rate of change in investment taking place in one of the n work/non-work segments i yields $\ln E_t = \ln E_0 + r_s S + r_p \sum_{i=1}^{n} \int_0^{e_i} (a_i + b_i t) \mathrm{d}t$. For the three-period case ($n = 3$), the earnings function is a quadratic in each work/non-work segment:

$$\ln E_t = \ln E_0 + r_s S + r_p \left(a_1 e_1 + \frac{1}{2} b_1 e_1^2 + a_2 e_2 + \frac{1}{2} b_2 e_2^2 + a_3 e_3 + \frac{1}{2} b_3 e_3^2 \right)$$

Taking a linear approximation of the quadratic in each segment and denoting segment e_2 as h (since it represents time at home out of the labor force) yields (6).

Jacobsen and Levin (1995), Kim and Polachek (1994), Light and Ureta (1995), Mincer and Ofek (1982), Mincer and Polachek (1974), Phipps et al. (2001), Rummery (1992), Sandell and Shapiro (1980), Sen (2001), Stafford and Sundstrom (1996), Stratton (1995), and Spivey (2005).

Intermittent Labor Force Participation and Human Capital Investment

As already illustrated, the lower the expected lifetime work, the smaller the gains from human capital investment, and the lower the amount invested. For this reason, a worker who anticipates discontinuous labor force participation procures less on-the-job training than the continuously employed worker. As a result, women's earnings need not exhibit the typical concave age-earnings profile characteristic of men. Instead, they are flatter and often exhibit a non-monotonic pattern depending on the degree of intermittent work behavior.

To see this analytically modify the life-cycle optimization model spelled out in Eqs. (1)–(3) above by introducing the possibility that labor force participation can vary year-by-year over the life cycle (Polachek 1975a). As such, modify (1) so that

$$Y(t) = R[N(t)E(t) - K(t)]$$

where $N(t)$ is the proportion of time available spent working in the labor force and investing in human capital. Assume $N(t)$ is exogenous to the investment process, but dependent on gender, marital status, and the number of children. Allowing for such intermittent labor force participation implies $N(t)$ is not constant in each period. This yields the following marginal gain from investment[21]:

$$\dot{\psi}(t) = -w_0 N(t) e^{r(t-T)} + w_0 r e^{rt} \int_t^T [N(\tau) - N(t)] d\tau$$

where again w_0 is the rental rate per unit of human capital, r the discount rate, and t one's current age.

The first term represents the marginal revenue if labor force participation was constant each time period. It is negative and identical to Ben-Porath's declining marginal gain from investment over the life cycle. The second term represents the incremental change to marginal revenue when labor force

[21] See Polachek (1975a) for a derivation.

participation is *not* constant over the life cycle. This term is positive if future labor force participation is expected to rise, as in the case when a woman anticipates reentering the labor force after raising her children. A sufficiently large second term implies an increasing present value of human capital investment. This means that intermittent labor force participation can cause human capital investment to rise during and after one's childrearing years instead of falling monotonically as Ben-Porath predicted. As such, post-school human capital investment (on-the-job training) crucially depends on expected lifetime labor force participation.

Most current studies of the gender wage gap do not take account of expected future labor force participation. As such, most overestimate the amount of the unexplained wage gap. However, one set of studies accounts for these biases. Polachek (1975a), Goldin and Polachek (1987), and Kao et al. (1994) analyze wage differences for the USA and for Taiwan. In contrast to traditional decomposition studies which explain 30–50% of the gender wage gap, these results explain up to 93% of the gap. To illustrate the robustness of the procedure, the technique used in these studies was applied within each gender to account for marital status wage differentials, specifically that married men earn more than single men, but married women earn less than single women. Here, these studies explain 82% of the marital status wage gap within each gender. Thus, lifetime work, governed by gender, marital status, and children, affects human capital acquisition, which in turn explains both why there is a gender wage gap and why there is a marital status gap that is opposite in magnitude for men and women. Whereas the human capital model emphasizes expected lifetime labor force participation, other studies also look at willingness to work long hours (Goldin 2014; Cortés and Pan 2016), workplace preferences (Wiswall and Zafar 2016), as well as psychological and motivational differences. These include payment scheme preferences (Dohmen and Falk 2001), time preference (Brown and van der Pol 2015), non-cognitive skills (Cobb-Clark and Tan 2011), mortality risk (Hammitt and Tuncel 2015), and risk preference (Booth and Katic 2013; Rai and Kimmel 2015). A survey of such articles is contained in Croson and Gneezy (2009).

Gender Wage Gap: Whites vs. Blacks

Related to lifetime work is the gender pay gap between whites and blacks. As indicated in Table 1, the gender earnings gap for blacks is smaller than for whites. One reason is lifetime labor force participation. At least in the past,

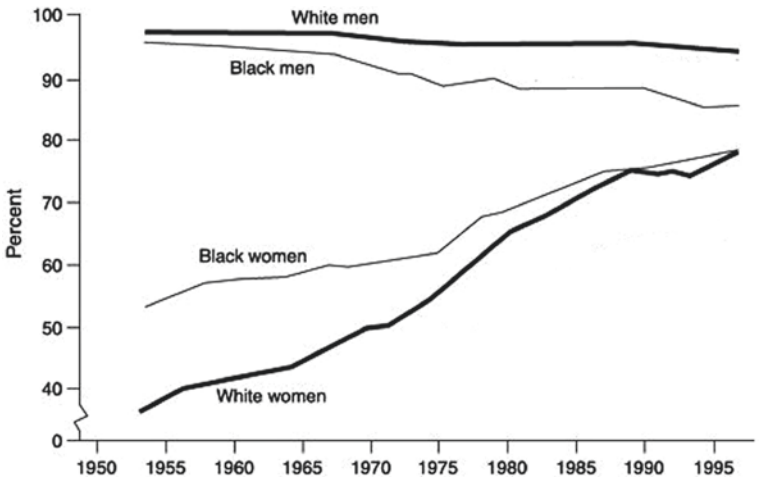

Fig. 4 Labor force participation by gender and race (*Source* https://www.dol.gov/wb/stats/facts_over_time.htm#labor)

black women worked slightly more over their lifetimes than white women, but black men compared to white men worked less. Figure 4 indicates racial differences in labor force participation. Although the data constitute annual rates, the figure is indicative of lifetime trends. The female–male earnings ratio for whites in 2015 is 0.78, but for blacks it is 0.90.[22]

Changes in Lifetime Labor Force Participation and the Gender Wage Gap

Changes in lifetime labor force participation can answer the second question, why the gender wage gap narrowed. At least since the time data have been collected, women's, especially married women's, labor force participation has risen. In 1890, only 4.9% of US married women participated. In 1948, this figure was approximately 33%, and in 2015, it was 57%. Figure 5 illustrates these labor force participation rates from 1948 to 2015. Higher labor force participation raises expected lifetime work and as a result increases human capital investments and wages. At the same time, male labor force participation declined moderately from 86% in 1948 to 70% in 2015. As such, female human capital investments most likely rose relative to males' human

[22]Based on data from: https://www.dol.gov/wb/resources/Womens_Earnings_and_the_Wage_Gap_17.pdf.

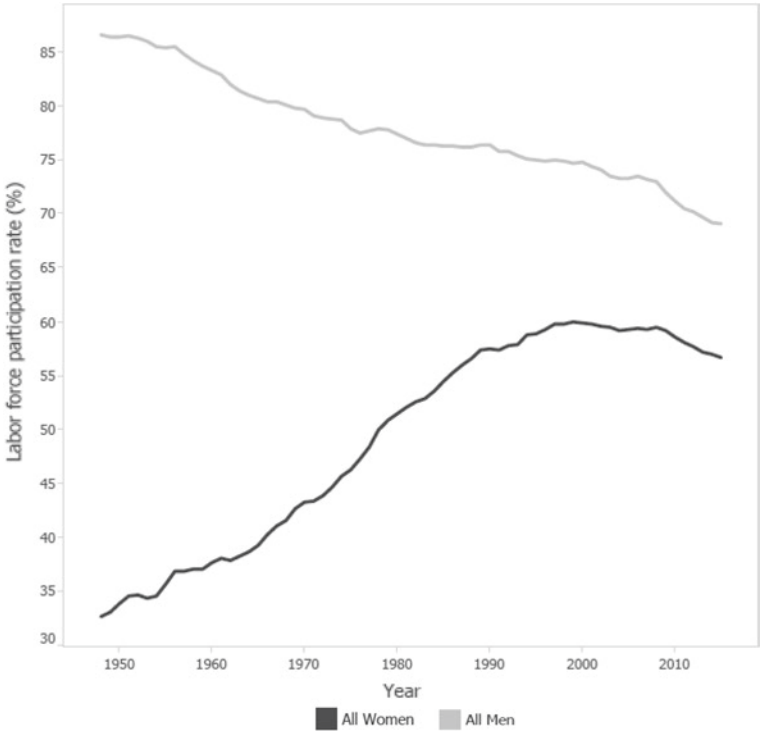

Fig. 5 Labor force participation rate by gender (1948–2015 annual averages) (*Notes* Includes persons in the civilian noninstitutional population that are employed or actively looking for work. Based on persons 16 years of age and older. *Source* 1948–2015 annual averages, Current Population Survey, US Bureau of Labor Statistics)

capital investments, thereby resulting in a higher female-to-male wage ratio. This is precisely what is observed in Fig. 6. However, there are exceptions, such as between 1960 and 1975. Polachek and Robst (2001, p. 869) found that the rapid rise in "new female labor force entrants in the 1970s brought down mean female wages, thereby driving down female wage growth." This is probably the case for the 1940s, as well, which witnessed an unprecedented influx of women workers during World War II.

Whereas the Mincer earnings function can be used to explain these demographic patterns of the earnings data, this formulation is insufficient with regard to other theoretical implications. For example, one shortcoming is that it does not explain how much schooling individuals actually accumulate in the first place. Indeed, within his model, individuals are indifferent between various amounts of school because each amount yields the same present value of lifetime earnings, thus making years of school an exogenous

Fig. 6 Gender earnings ratio (March 1960–2014) (*Source* US Bureau of Labor Statistics)

variable. But as is well known from the life-cycle human capital model, an individual chooses his or her years of schooling based on the five parameters (b, β, E_0, δ, and r) alluded to earlier in this chapter. Thus, one must estimate these five structural parameters person by person to test the theory's implications. Nowadays, sufficiently long panel data are available to follow each person for a long enough time period to obtain person-specific estimates. We do so now.

6 Human Capital Production Function Parameter Values

Among the first to estimate nonlinear (in the parameters) earnings functions was Haley (1976). He used CPS (Series P-60, No. 56) schooling and earnings (unfortunately earned and unearned income) data for individuals 18–64 in 1956, 1958, 1961, 1963, 1964, and 1966, thus implying the pooling of 6 cross sections. However, his slightly more complex formulation had identification problems precluding his capability to estimate $E_0, \beta,$ and R.

Nevertheless, the crucial parameters $b, r,$ and δ were obtained for seven schooling levels, along with parameters defining earnings growth across cohorts. Most of Haley's estimates are as expected. For example, discount rate (r) estimates are between 5 and 7%, the b ability coefficient is about 0.6, and depreciation (δ) is between 0.019 and 0.043. These estimates compare favorably to other studies that estimate aggregate Ben-Porath-based models, though understandably there are differences due to alternative methodologies and data. More specifically, with respect to the Cobb-Douglas production function exponent b, Heckman's (1975) 0.67, Heckman's (1976) 0.51–0.54, Heckman et al.'s (1998) 0.80, Song and Jones's (2006) 0.5, and Liu's (2009) 0.52 span Haley's 0.54–0.60. For depreciation (δ), Johnson and Hebein's (1974) 0.022 and Heckman's (1976) 0.04–0.07 estimates compare favorably to Haley's 0.017–0.043. Similarly, Rosen's (1976) 0.07 discount rate (r) estimate is Haley's upper bound 0.07. Further, as already mentioned, each uses slightly different human capital production functions, and some incorporate life-cycle labor supply. On the other hand, not all the human capital theory's predictions are observed in Haley's estimates. For one, a higher b should imply more schooling, but Haley does not find this. Also, the relationship between skill depreciation and schooling level should be negative, but this is not the case in Haley's empirical work. One weakness in these studies is not adequately taking account of heterogeneity.

6.1 Heterogeneity

The advent of speedier computers, better optimization routines, and longer panels than in the past now enables one to retrieve individual-specific parameters of the human capital life-cycle model by estimating appropriate earnings functions individual-by-individual. This allows one to account for heterogeneity because ability-type parameters can be estimated for each person. The first to do this is PDT (2015). They obtain the five parameters $b, r, \delta, E_0,$ and β, for about 1700 male workers contained in the National Longitudinal Survey of Youth (NLS-Y), as well as a population-wide value for the rental rate of human capital R. They plot kernel density functions and find significant heterogeneity. Important to macroeconomists, accounting for this heterogeneity dramatically reduces estimates of population-wide persistence of permanent and transitory shocks in earnings dynamics models by over 50%.[23]

[23]Other studies concentrate on heterogeneity by allowing ARMA processes to vary across individuals (e.g., Browning and Ejrnæs 2013). Some present decile ranges of key parameters illustrating that

PDT's technique also yields a number of additional new findings. For example, on the microlevel, they find that blacks have higher rates of skill depreciation than whites. Here in this chapter, we extend PDT's work to include Hispanics and present average coefficient estimates for them as well as for blacks and whites in Table 6. More interestingly, whereas typical ability measures obtained from Armed Forces Qualification Test (AFQT) test scores[24] differ widely between Hispanics, blacks, and whites, there are far smaller differences in the human capital ability parameters $b, \beta,$ and E_0 indicating possible racial biases in typical psychology-based aptitude, achievement, and intelligence test scores. This is consistent with Fryer and Levitt (2004, 2013) who find small racial differences in IQ once adjusting for a number of demographic factors.

6.2 Implications Regarding Individual-Specific Human Capital Parameters

Schooling Levels and Human Capital Parameters

Obtaining individual-specific parameters enabled PDT to verify a number of previously untested theorems based on the life-cycle human capital model. More specifically, human capital theory predicts a positive correlation between ability measures b and β and one's years of schooling, a negative relation between initial stock of human capital E_0 and schooling level, and negative correlations between a person's years of school and his/her discount (r) and skill depreciation (δ) rates. Greater ability to learn raises the amount of human capital one can produce per unit of time, thus lowering the cost of human capital acquisition and increasing the amount of school obtained. On the other hand, more initial human capital E_0 is a substitute for schooling and thus leads one to stop school earlier. Higher depreciation rates lower the amount of human capital retained, thus making school relatively more costly and decreasing the amount purchased. Finally, schooling levels decrease with time discount (r) because individuals with high discount rates are more reluctant to put off the gratification of current market earnings.

---

heterogeneity affects the speed individuals respond to shocks (e.g., Browning et al. 2010; Browning and Ejrnæs 2013). In other realms, Greene (2005, 2010) examines heterogeneity by using fixed- and random-effects models.

[24]AFQT scores are computed using the Standard Scores from four ASVAB subtests: Arithmetic Reasoning (AR), Mathematics Knowledge (MK), Paragraph Comprehension (PC), and Word Knowledge (WK).

Table 6 Mean and standard deviation of the parameter estimate (by race)

	Mean	SD
Hispanic (436 persons)		
b	0.33	0.09
β	0.61	0.18
E_0	2.95	3.42
δ	0.028	0.017
r	0.044	0.041
Average weekly earnings (1982–1984$)	354	258
t	31.58	8.48
t^*	17.07	2.39
AFQT	30.61	26.10
Black (596 persons)		
b	0.32	0.12
β	0.57	0.16
E_0	2.73	3.41
δ	0.029	0.016
r	0.043	0.042
Average weekly earnings (1982–1984$)	309	243
t	31.94	8.48
t^*	17.71	1.87
AFQT	20.41	19.49
White (1230 persons)		
b	0.36	0.09
β	0.65	0.17
E_0	2.76	2.69
δ	0.026	0.014
r	0.041	0.038
Average weekly earnings (1982–1984$)	443	358
t	32.05	8.48
t^*	18.18	2.22
AFQT	52.35	27.78

Source Based on the data in Polachek et al. (2015)

Personality and Human Capital Parameters

Similarly, PDT find human capital parameters to be related to personality. They observe greater ability as well as lower skill depreciation and time discount rates for those individuals with a high internal locus of control and for those individuals who demonstrate high levels of self-esteem. Individuals inclined toward mental depression have a higher time discount. At the same time, family background, such as higher parental education, is associated with a greater ability to learn, lower skill depreciation, and a smaller rate of time discount. Educational stimuli, such as growing up in a household that subscribed to newspapers and magazines, are associated with

Table 7 Correlation among estimated parameters and standardized test scored (by race)

	All					Hispanics					Black					Whites				
	b	β	E_0	δ	r	b	β	E_0	δ	r	b	β	E_0	δ	r	b	β	E_0	δ	r
Cognitive																				
Gen Sc.	0.19	0.18	0.05	−0.12	−0.03	0.17	0.24	0.04	−0.01	0.00	0.09	0.26	−0.01	−0.05	−0.06	0.17	0.10	0.07	−0.12	−0.02
Arithmetic	0.22	0.20	0.04	−0.13	−0.05	0.22	0.21	0.01	−0.03	−0.07	0.07	0.22	0.01	−0.08	−0.07	0.22	0.15	0.06	−0.13	−0.04
Word know	0.19	0.19	0.02	−0.14	−0.05	0.12	0.21	0.05	−0.04	0.03	0.08	0.20	−0.07	−0.09	−0.10	0.17	0.12	0.04	−0.13	−0.04
Para Comp	0.16	0.17	0.03	−0.14	−0.03	0.11	0.20	0.00	−0.07	0.01	0.07	0.18	−0.08	−0.13	−0.14	0.14	0.11	0.06	−0.14	−0.01
Numeric	0.21	0.26	0.05	−0.11	−0.04	0.19	0.26	0.07	−0.05	−0.02	0.08	0.25	−0.02	−0.06	−0.05	0.21	0.22	0.07	−0.11	−0.04
Coding	0.19	0.20	0.04	−0.12	−0.06	0.18	0.25	0.07	−0.02	−0.02	0.04	0.24	−0.03	−0.06	−0.07	0.18	0.15	0.05	−0.12	−0.06
Auto	0.07	0.19	0.06	−0.09	0.04	0.10	0.21	0.00	−0.02	0.10	0.03	0.18	0.02	−0.06	0.01	0.01	0.12	0.09	−0.07	0.05
Math know	0.23	0.21	0.02	−0.13	−0.10	0.21	0.14	−0.01	−0.09	−0.11	0.10	0.25	−0.04	−0.06	−0.12	0.23	0.17	0.03	−0.13	−0.10
Mechanical	0.15	0.17	0.04	−0.12	−0.01	0.13	0.21	0.02	0.02	0.00	0.06	0.16	−0.03	−0.07	−0.08	0.12	0.11	0.06	−0.12	0.01
Electronics	0.14	0.20	0.07	−0.11	0.02	0.12	0.21	0.05	−0.02	0.04	0.05	0.23	0.03	−0.05	−0.03	0.11	0.14	0.10	−0.11	0.04
AFQT (raw)	0.22	0.21	0.03	−0.15	−0.06	0.17	0.22	0.03	−0.06	−0.02	0.09	0.23	−0.07	−0.11	−0.12	0.21	0.15	0.05	−0.15	−0.05
Non-cognitive																				
Rotter	−0.12	−0.08	0.00	0.11	0.04	−0.02	−0.08	−0.10	−0.02	−0.05	−0.16	−0.06	−0.02	0.05	0.05	−0.11	−0.07	0.01	0.13	0.05
Self estm80	0.08	0.13	0.08	−0.08	0.06	0.13	0.25	0.12	0.03	0.03	0.05	0.14	0.06	−0.08	−0.02	0.07	0.11	0.09	−0.08	0.08
Pearlin	0.13	0.13	0.00	−0.10	−0.11	0.16	0.18	0.07	0.01	−0.08	0.11	0.14	−0.07	−0.06	−0.09	0.13	0.11	0.01	−0.11	−0.11
Trust	−0.11	−0.08	0.00	0.08	0.04	−0.11	−0.02	−0.03	0.06	0.04	−0.04	−0.09	0.04	0.01	0.13	−0.09	−0.06	0.00	0.08	0.02
CESD20	−0.10	−0.08	0.03	0.10	0.05	−0.07	−0.17	0.06	0.01	0.04	−0.17	−0.03	0.06	0.11	0.07	−0.07	−0.05	0.02	0.10	0.04
Family background																				
Mother edu	0.16	0.08	−0.04	−0.11	−0.07	0.17	0.10	−0.03	−0.10	−0.06	0.03	0.20	0.10	0.00	0.02	0.17	0.02	−0.07	−0.13	−0.09
Father edu	0.16	0.13	−0.04	−0.09	−0.10	0.22	0.07	0.00	−0.08	−0.08	0.02	0.12	0.06	0.01	−0.05	0.15	0.10	−0.06	−0.10	−0.11
Urban	−0.02	−0.07	−0.02	−0.02	−0.01	−0.11	−0.05	0.03	0.02	0.02	0.05	−0.04	−0.08	0.01	−0.02	−0.04	−0.09	−0.02	−0.02	−0.01
Magazine	0.15	0.11	0.01	−0.12	−0.06	0.11	0.13	0.02	0.01	−0.02	0.07	0.10	−0.01	−0.13	−0.04	0.14	0.06	0.02	−0.12	−0.06
Newspaper	0.08	0.11	0.04	−0.04	0.00	0.07	0.09	0.03	−0.04	−0.03	−0.04	0.13	0.05	−0.05	0.03	0.08	0.06	0.04	−0.02	0.01
Library	0.10	0.06	−0.02	−0.05	−0.05	0.06	0.21	0.01	0.03	0.04	0.06	0.05	0.11	−0.02	−0.03	0.10	0.03	−0.05	−0.05	−0.06
Poverty	−0.09	−0.15	−0.04	0.04	−0.03	−0.13	−0.13	−0.01	0.07	−0.05	−0.05	−0.19	−0.16	−0.05	−0.08	−0.05	−0.08	−0.01	0.04	−0.04

Note AFQT represents Armed Force Qualification Test; CESD represents 20 question depression index
Source Polachek et al. (2015)

a higher ability. Conversely, growing up poor is associated with lower levels of ability. These correlations which are now broken down by race and ethnicity are given in Table 7.

6.3 Homogeneity vs. Heterogeneity of Human Capital

Strictly speaking, the human capital model assumes potential earnings are directly related to the amount of human capital one purchases throughout one's life ($Y_t^P = RE_t$) and observed earnings equal potential earnings minus current investments $Y_t = R[E_t - K_t]$. Observed earnings rise with age as one accumulates more human capital, but eventually fall when skill depreciation outweighs new investments in human capital. Underlying this approach is an assumption that all human capital is homogeneous because everyone faces the same rental rate per unit of human capital. One earns more because one has accumulated more human capital, but not because one has a different *type* of human capital. But it is not obvious that human capital is homogeneous, and thus, it is not obvious that all earnings variations come about because *amounts* and not *types* of human capital differ from person to person across the population. For example, holding years of school constant, do newly graduating engineers earn more than new humanities majors because engineers have more human capital, or do new engineers earn more because they bought a different type human capital? In other words, is human capital homogeneous or is it heterogeneous?

A number of papers claim the latter. For example, Polachek (1979, 1981) argues in favor of heterogeneity. He devises a matching model in which the production function for human capital varies by occupation. Although many human capital production function parameters can vary, because of his interest in gender occupational segregation he concentrates simply on skill depreciation due to non-use of human capital (atrophy) when dropping out of the labor force. As such, he assumes $\dot{E} = f(K_t) - [\delta + (1 - N_t)\xi]E_t$ where ξ is an occupation-specific atrophy rate and N_t is the proportion of time working in year t.[25] Given that compensating market differentials likely rewards high-depreciation occupations more generously, the human capital rental rate (R) should increase with atrophy, implying $R = R(\xi)$ such that $R'(\xi) > 0$. He shows that those individuals more likely to drop out

[25]Atrophy is zero when N_t is 1, but is ξE_t when N_t is 0.

will plausibly choose occupations with low atrophy rates. Based on this, he explains a large amount of gender-based occupational differences.[26]

One aspect of the PDT identification strategy is their approach to measure *R*, the population-wide human capital rental rate.[27] PDT can do this because the human capital model assumes everyone faces a common market-wide rental rate *R* determined solely by aggregate market forces. In contrast, the parameters governing the production of human capital vary by individual based on each person's ability. But one can go farther by determining whether *R* is indeed constant across the population. Checking this hypothesis gets at a direct test of human capital homogeneity. Homogeneity implies that each basic human capital unit rents for the common price determined in the market. Under homogeneity, this rental rate should be the same, independent of any factor, since human capital in all endeavors is comparable. However, heterogeneity implies rental rates can differ if the market rewards various types of human capital dissimilarly. In short, human capital is homogeneous if rental rates remain constant, but is heterogeneous if rental rates vary by type of human capital. Obviously, nonmarket considerations such as discrimination, regional variations, or time-varying macroeconomic conditions can tweak the rental rate, since supply and demand fluctuations can alter spot market prices.

PDT test for homogeneity. They find very little variation in rental rates across industries, across broad occupations, or across schooling levels. Only unemployment is negatively correlated with rental rates, which makes sense since a weak economy lowers aggregate demand, but they also find slight race differences, perhaps getting at discrimination. Preliminary research by Andrew Verdon (2018) corroborates this for the UK (using the British Household Panel Survey) and Korea (using the Korean Labor and Income Panel Study), but finds rental rate differs by industry in Germany (using German Socio-Economic Panel) and by occupation in the USA using PSID data, though more research on this is needed.

[26]Heckman, Layne-Farrar, and Todd (1996) also claim heterogeneity in human capital. They do so by exploiting three interactions: (1) between school quality and education, (2) between regional labor shocks and education, and (3) between place of birth and place of residence.

[27]Heckman et al. (1998) adopt an alternative identification strategy to determine *R*. Their approach exploits the fact that all observed earnings changes (adjusted for hours) between two time periods must be attributed to rental rates changes when in "flat periods" a time when human capital stock (E_t) remains constant. Typically, flat spots occur late in life, usually around the mid-fifties, an age greater than any current respondent in the NLSY. Bowlus and Robinson (2012), who apply the flat spot identification approach with CPS data, obtain similar results to PDT.

One limitation of PDT is they perform the analysis only for men. As stated earlier, a structural earnings equation for potentially discontinuous workers is far more complex and less tractable empirically.

7 Inequality: Comparing the Impact of Observables and Unobservables

Schooling and experience are the key to easily observable workhorse variables used in past human capital studies to explain individual variations in earnings. However, PDT show that a person's previously unobservable abilities (b, β, and E_0) and a person's time preference (r) and skill depreciation (δ) are also important in explaining earnings differences. To get at the relative contributions of these new individual-specific b, β, E_0, r, and δ parameters, we compare their importance to the importance of school and experience in determining earnings levels and earnings disparities (variance). We do this in three ways. First, we compute elasticities, namely the percent impact on earnings of a given percent change in b, β, E_0, r, and δ, and compare these elasticities to comparably computed school and experience earnings elasticities.[28] Second, we compute the impact of b, β, E_0, r, and δ on their overall explanatory power (R^2) in determining earnings and compare these to the explanatory power of school and experience. Finally third, we compute the impact of b, β, E_0, r, and δ on a measure of earnings distribution (σ_y^2) and compare these to comparable measures for school and experience.

7.1 Earnings Levels

One way to determine the relative importance of schooling and experience compared to previously unobserved ability $b, \beta, \text{and } E_0$ is to compute the impact of a given percent increase in each set of variables on the percent change in earnings that results. Such "elasticities" can be obtained based on Appendix Eq. (9) by examining the degree to which earnings rise when increasing school and experience by a given percent compared to the percent rise in earnings when one increases $b, \beta, \text{and } E_0$ by the same percent. We report results for this exercise in Table 8.

[28]Another similar approach is to compute the percent impact on earnings of a standard deviation increase in each variable.

Table 8 Earnings elasticities with respect to structural parameters, age, and school leaving age (t^*)

	b	β	E_0	δ	r	t	t^*
Hispanics	1.00	1.60	0.16	−0.60	−0.03	1.17	0.19
Blacks	1.00	2.06	0.20	−0.86	−0.03	1.42	0.23
Whites	1.33	1.98	0.19	−0.82	−0.04	1.39	0.24
All	1.27	1.97	0.19	−0.81	−0.04	1.38	0.24

Note Computations are based on the earnings function and data given in Polachek et al. (2015)

More specifically, the table shows that, on average, a 10% rise in b and β leads to a 12.7 and 19.7% rise in earnings. A 10% increase in experience (t) yields a 13.8% rise in earnings, but interestingly a 10% increase in years of school only augments earnings 2.4%. As expected, higher δ values reflect skill deterioration thereby leading to lower earnings. Thus, a 10% rise in δ leads to an 8.1% decline in earnings. Also, relatively of small importance are E_0 and r. The elasticity of earnings with respect to E_0 is 0.19 and with respect to r is −0.04. In summary, b, β, t, and δ are relatively important, whereas schooling, E_0, and r are not.

The earnings elasticities with respect to b and schooling are slightly higher for whites, whereby the earnings elasticities with respect to β and E_0 are slightly higher for blacks. Earnings elasticities for Hispanics are lower than blacks and whites.

Table 8 presents average elasticities indicating the impact observable and previously unobservable attributes have on earnings levels. However, the effects can be nonlinear. We use individual-specific parameters based on PDT to get at this nonlinearity. We plot these nonlinear elasticities over the range of parameter values. Figure 7 contains these relationships.

The elasticity with respect to b increases as the value of b rises. This means that an intervention that raises b will increase earnings by a greater percent for those already with a high b. In short, the more able will benefit more.

The pattern for β is the opposite. These elasticities decrease with β. This means that an intervention that increases β will increase earnings proportionally more for those individuals with lower β.

The earnings elasticities with respect to E_0 and schooling have similar patterns to each other. They first rise as the level of E_0 and schooling rise, and then decline. These similar patterns are expected as both schooling and E_0 represent stocks of human capital. With regard to schooling, the inverted U-shape indicates an increasing importance of school up until post-baccalaureate education.

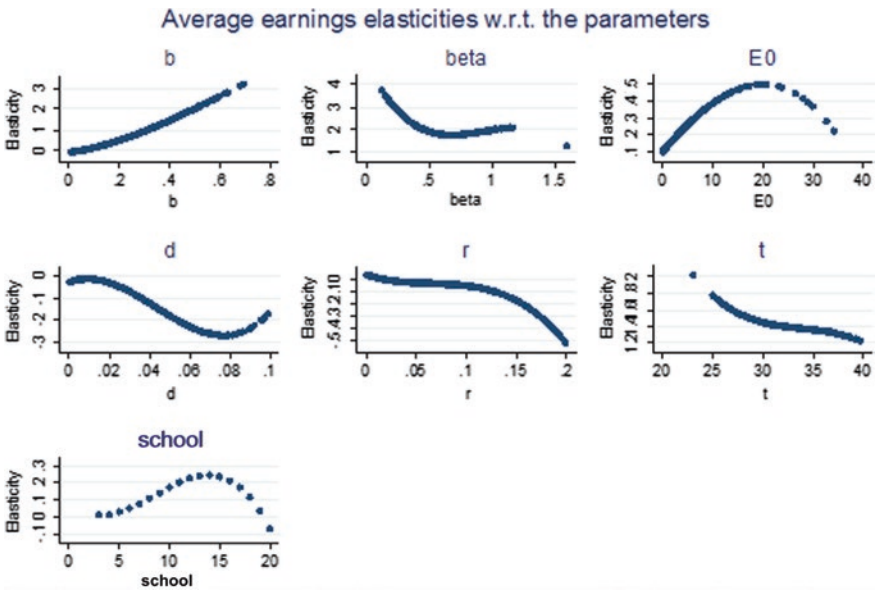

Fig. 7 Earnings elasticities with respect to personal attributes (*Notes* Graphs represent predicted elasticities obtained from cubic regressions. *Source* PDT [2015]; our computations)

The effect of the skill depreciation rate (δ) and time discount rate (r) on elasticities is somewhat similar. As δ and r rise, the elasticities decline at an increasing rate. Only at a very high level of δ, there is a slight upward trend, but the magnitude of the elasticities still remains negative.

7.2 Explanatory Power

The relative importance of unobserved ability b, β, E_0, and depreciation and discount rates δ and r compared to schooling and experience is to compute the contribution to R^2 of each of these factors while holding the other factors constant. To do this, we successively run regressions of the form

$$\ln y = \alpha_0 + \alpha_1 b + \alpha_2 \beta + \alpha_3 E_0 + \alpha_4 r + \alpha_5 \delta + \alpha_6 S + \alpha_7 t + \epsilon$$

in which we vary only one of the b_i, β_i, r_i, δ_i, E_{0i},, schooling, and t variables while holding the others constant at their mean levels. This exercise leads to seven regressions (for each age group). The various R^2 values are presented in Table 9.

First, the explanatory power of each of the attributes is fairly stable across each age group. Second, the parameter β has the highest explanatory power.

Table 9 Explaining earnings variance, ability, time preference, age, and school leaving age

Age group	b	β	E_0	δ	r	t	t^*
20–24							
Obs	8408	8408	8408	8408	8408	8408	8408
R^2	0.148	0.288	0.018	0.167	0.003	0.002	0.002
25–29							
Obs	10,728	10,728	10,728	10,728	10,728	10,728	10,728
R^2	0.155	0.282	0.018	0.160	0.003	0.001	0.003
30–34							
Obs	8640	8640	8640	8640	8640	8640	8640
R^2	0.157	0.281	0.019	0.162	0.003	0.001	0.004
35–39							
Obs	5465	5465	5465	5465	5465	5465	5465
R^2	0.155	0.282	0.018	0.158	0.003	0.001	0.004
40–44							
Obs	4879	4879	4879	4879	4879	4879	4879
R^2	0.161	0.280	0.019	0.160	0.003	0.001	0.004
45–49							
Obs	3853	3853	3853	3853	3853	3853	3853
R^2	0.155	0.288	0.019	0.152	0.003	0.001	0.004
50–54							
Obs	897	897	897	897	897	897	897
R^2	0.218	0.299	0.023	0.156	0.003	0.000	0.005

Note R^2 is computed as the ratio of variance of the predicted earnings based on each factor to the variance of the actual earnings. Predicted earnings for each factor are calculated by allowing that factor to vary, while holding all other factors constant
Source Computed based on the data and earnings function from Polachek et al. (2015)

Third, the explanatory power of the b and β abilities to learn and the human capital depreciation rate are substantially higher than the explanatory power of E_0 and schooling. Fourth, time preference plays almost a negligible role in explaining the earnings variance. And, fifth, in absolute terms, schooling and E_0 have very little explanatory power.

Noteworthy is the observed weak explanatory power of schooling. In a sense, this is paradoxical, especially since most past studies argue that school is the most important determinant of earnings. Yet we find schooling to play a more minor role compared to b, β, and δ. These three parameters alone respectively reflect the highest relative strength in explaining earnings variation. Thus, the results imply that ability is more important in determining earnings than school level per se. Not only does one's ability dictate one's schooling level, but also a higher ability enables one to produce more human capital while in school. Further, skill depreciation (δ) which indicates the degree one retains knowledge also contributes greatly. Thus, the ability to

Table 10 Linear and log-linear earnings regressions

	Without school in regression	With school in regression	
	Adjusted R^2	Adjusted R^2	Coefficient (Sch)
Dependent variable: y			
exp, exp2	0.100	0.268	65.95
exp, exp2, AFQT	0.219	0.287	51.37
exp, exp2, b, β, δ, E_0, r	0.460	0.500	35.08
exp, exp2, b, β, δ, E_0, r, AFQT	0.469	0.499	35.13
Dependent variable: log(y)			
exp, exp2	0.14	0.292	0.134
exp, exp2, AFQT	0.274	0.324	0.093
exp, exp2, b, β, δ, E_0, r	0.566	0.596	0.065
exp, exp2, b, β, δ, E_0, r, AFQT	0.577	0.597	0.061

Note Data obtained from NLSY79 and PDT (2015)
Source NLSY79; Polachek et al. (2015)

learn and retain knowledge seems to be the important determinants of earnings. In a sense, this finding is consistent with work to be discussed shortly on how past studies overestimate schooling rates of return by inappropriately accounting for ability.

Another way to look at this is to examine the contribution to R^2 in a typical Mincer earnings function (our Eq. (2)). Table 10 reports adjusted R^2 measures for various specifications of the earnings function. AFQT increases the adjusted R^2 by only 0.04 over schooling and experience in a linear fit, whereas b, β, and E_0 increase adjusted R^2 by 0.19. Incorporating AFQT adds virtually nothing (0.01) when including PDT's other three ability measures b, β, and E_0. Adding schooling (Column 2) raises the explanatory power only when ability is not included. AFQT essentially does nothing when b, β, r, δ, and E_0 are already in the regression. Thus, the five human capital parameters jointly explain earnings more than schooling and traditionally measured ability (AFQT).

Bowles et al. (2001) examine the role of ability by estimating the rate of return to schooling before and after adjusting for cognitive test scores. Essentially, they compare the schooling coefficient for the "years-in-school" variable in our Eq. (2) with the schooling coefficient in an equation when a cognitive test score is introduced as an independent variable (neglecting endogeneity issues). They find cognitive ability reduces the schooling coefficient by 18% on average and thus conclude "a substantial portion of the returns to schooling are generated by effects or correlates of schooling substantially unrelated to the cognitive capacities measured on the available tests." Our replication using the

AFQT test for individuals in the 1979 NLS-Y yields 31%.[29] Replicating this using our five ability parameters yields a reduction of 51%.[30] Further, adding AFQT has only minimal effect. Thus, we conclude ability matters, but usual ability measures obtained via typical psychometric tests, such as the AFQT, do not get at all facets of ability, particularly they do not get at the type of abilities that matter with respect to real-world accomplishments.

7.3 Variance Decomposition

Our third approach is to decompose the earnings variance into that part attributable to observable schooling and experience, and that part attributable to b, β, r, δ, and E_0. Chiswick and Mincer (1972) devise a framework to identify sources of earnings inequality. Their approach concentrates on schooling and work experience which they find explain a substantial portion of the earnings inequality. However, they cannot evaluate the role differences in individual abilities, time discount, and skill depreciation rates (b_i, β_i, r_i, δ_i, and E_{0_i}) play because they do not estimate individual-specific parameters. However, based on PDT's individual-specific estimates and the structure of the human capital framework, we can assess the relative importance of these parameters. We examine the sensitivity of earnings variance to changes in the variation in these factors.

To answer this question, we conduct a variance decomposition exercise. Unlike in Chiswick and Mincer (1972), the earnings function we use is the nonlinear function given in PDT. The complex nonlinearity makes variance decomposition difficult. To circumvent this difficulty, we first linearize it with a first-order Taylor series expansion and then conduct the variance decomposition on the linearized version:

$$\begin{aligned} f(b,\beta,E_0,\delta,r) \approx f\left(b_a,\beta_a,E_{0a},\delta_a,r_a,t,t^*\right) &+ f_b^a(.)(b-b_a) \\ &+ f_\beta^a(.)(\beta-\beta_a) + f_{E_0}^a(.)(E_0-E_{0a}) + f_\delta^a(.)(\delta-\delta_a) \\ &+ f_r^a(.)(r-r_a) + f_t^a(.)(t-t_a) + f_{t^*}^a(.)\left(t^*-t_a^*\right) \end{aligned}$$

where $b_a, \beta_a, E_{0a}, \delta_a, r_a, t_a, t_a^*$ are the average of $b, \beta, E_0, \delta, r, t, t^*$, $f^a(.)$s are the corresponding partial derivatives of the earnings function with

[29]Computed as 1-(0.093/0.134) from row (2) of column (3) in the lower panel of Table 10.

[30]Computed as 1-(0.065/0.134) from row (3) of column (3) in the lower panel of Table 10.

respect to each of the factors respectively and evaluated at the mean values of $b, \beta, E_0, \delta, r, t, t^*$. Collecting terms and adding an error ϵ yield

$$Y \approx A + f_b^a(.)b + f_\beta^a(.)\beta + f_{E_0}^a(.)E_0 + f_\delta^a(.)\delta + f_r^a(.)r + f_t^a(.)t + f_{t^*}^a(.)\left(t^*\right) + \epsilon.$$

Assuming $b, \beta, E_0, \delta, r, t, t^*$ are uncorrelated with ϵ, the variance of Y (i.e., σ_Y^2) in terms of the right-hand-side variables is

$$\begin{aligned} \sigma_Y^2 &= \sum_m f_m^2(.)\sigma_m^2 + \sum_{m \neq l} f_m(.)f_l(.)Cov(m,l) + \sigma_\epsilon^2 \\ &= \sum_m f_m^2(.)\sigma_m^2 + \sum_{m \neq l} f_m(.)f_l(.)\sigma_m\sigma_l R_{ml} + \sigma_\epsilon^2 \end{aligned} \tag{7}$$

where $m, l = b, \beta, E_0, \delta, r, t, t^*$, σ_m are the standard deviations, and R_{ml} are the pairwise correlation coefficients between m and l. Table 11 presents the values of each component of σ_Y^2.

Expression (7) enables one to assess the effect of a change in the standard deviation of a right-hand-side variable on the variance of earnings. Taking partial derivatives with respect to each of the factors yields the following:

$$\frac{\partial \sigma_Y^2}{\partial \sigma_m} = 2f_m^2(.)\sigma_m + 2f_m(.)\textstyle\sum_{m \neq l} f_l(.)\sigma_l R_{ml} \tag{8}$$

Multiplying both sides of (8) by (σ_m/σ_Y^2) gives the elasticity of σ_Y^2 with respect to σ_m. These elasticities for each of the factors are in Table 12.

The results suggest that for every 10% decline in standard deviation of b, the variance of earnings declines by 2.1%. The effect of a change in the standard deviation of β on the variance of earnings is slightly larger. The

Table 11 Components of σ_Y^2

			Correlation coefficients						
	$f^a(.)$	$SD(\sigma)$	b	β	E_0	δ	r	t	t^*
b	948.8	0.103	1	0.011	−0.152	−0.024	−0.247	0.025	0.217
β	735.2	0.172	0.011	1	0.363	0.475	0.047	0.016	0.145
E_0	14.7	3.040	−0.152	0.363	1	0.474	0.409	0.016	−0.041
δ	−6824.5	0.015	−0.024	0.475	0.474	1	0.134	−0.034	−0.169
r	−307.3	0.040	−0.247	0.047	0.409	0.134	1	0.005	−0.145
t	5.8	8.458	0.025	0.016	0.016	−0.034	0.005	1	0.131
$t*$	6.7	2.208	0.217	0.145	−0.041	−0.169	−0.145	0.131	1

Note Computations based on the data and earnings function given in Polachek et al. (2015)
Source Polachek et al. (2015); our computations

Table 12 Earnings variance elasticities (σ_Y^2)

b	β	E_0	δ	r	t	t^*
0.21	0.26	0.02	0.07	0.01	0.06	0.02

Note Coefficients are the percent impact on the variance of earnings of an increase in the variance of the indicated parameters
Source Polachek et al. (2015); our computations

elasticities with respect to the standard deviation of other parameters, t, and t^* are relatively small. This result again implies that one's ability to create new human capital from old is the most important factor determining earnings distribution. In short, ability matters.

8 Endogeneity Issues: Causal Effect Estimation

Over the past few decades, researchers have identified a number of factors and estimated their impact on earnings and the earnings distribution. A large number of identification strategies were proposed to establish causal effects. The basic idea underlying these methods is to generate exogenous variation in the explanatory variables so that the causal impacts are identified without other potential confounding factors. Earlier studies on this topic assume that independent variables are exogenous and apply OLS. However, as the potential biases originating from omission of relevant variables and non-representative sample selection were recognized, researchers adopted a variety of alternative identification strategies. These include instrumental variables, twin comparisons, and natural or quasi-natural experiments.

The most widely studied topic is the effect of years of schooling on earnings. A large number of papers appeared since the early 1990s that apply the instrumental variable method to estimate the return to schooling (Angrist and Krueger 1991; Ashenfelter and Krueger 1994; Ashenfelter and Rouse 1998; Kane and Rouse 1995; Card 1995, 1999; Harmon and Walker 1995; Staiger and Stock 1997; Conneely and Uusitalo 1997; Ichino and Winter-Ebmer 2004; Lemieux and Card 1998; Meghir and Palme 1999; Maluccio 1997; Duflo 2001). The estimates from these studies vary widely, ranging from 3.6 to 94.7% (Card 2001).

Despite the volume of the previous work, the validity of many of the IVs used so far remains unclear. Specifically, the exclusion restriction condition imposed on these IVs became the main point of concern. For instance, Card (1995) uses geographic proximity to college as an instrument in an earnings

regression. Presumably being near a college reduces the cost of attendance, for example, by allowing students to live at home. Thus living nearby increases college attendance but by itself is not correlated with other unobserved factors influencing earnings. However, this assertion received a mixed reaction. Carneiro and Heckman (2002) show that distance to college in the NLSY79 is correlated with ability thereby violating the exclusion restriction. Slichter (2015) also concludes that geographic propinquity to college is an invalid instrument and likely results in an overestimate of the returns to college education. On the other hand, Kitagawa (2015) finds no evidence of its invalidity as an instrument when also adjusting for race, region, job experience, parental education, and whether one lives in an urban area.

Another well-cited instrument is the quarter of birth used by Angrist and Krueger (1991). Students born at the beginning of the academic year are older. A good number of these leave school upon reaching the minimum compulsory dropout age, thus having one less year of school than their counterparts born slightly later. In essence, they use an estimate of the earnings impact of this extra year of school as an unbiased estimate of the return under the assumption birth quarter is random. Despite its appeal, Bound and Jaeger (1996) criticize this approach. They present a number of studies that show that quarter of birth may be an invalid instrument because it is correlated with other determinants of earnings. These include studies showing quarter of birth to be correlated with mental illness, retardation, personality, and family income. Further, a placebo test using data predating the compulsory school laws yields the same result that birth quarter affects earnings.

Another substantive concern with IV-based estimation is the use of weak instruments (Staiger and Stock 1997; Kleibergen 2002; Moreira 2003). For instance, Angrist and Krueger (1991) use a number of weak instruments as many of their first-stage F-statistics are less than 5 (Staiger and Stock 1997). Bound et al. (1995) argue that the use of a large number of weak instruments makes the IV estimates move closer to OLS. Using the same data as in Angrist and Krueger (1991), Bound et al. (1995) replace the quarter of birth IV by irrelevant random numbers and estimate 6% returns to schooling with an estimated standard error of $\pm 1.5\%$ (see Imbens and Rosenbaum 2005).

Due to these limitations, an alternative literature emerged that uses a partial identification strategy. The attractive feature of this approach is that it relies on weaker yet more credible assumptions than the ones necessary for standard IV-based regressions. However, the approach leads to a bounded estimate of the causal effect rather than a point estimate. Manski and Pepper

(2000, 2009) develop a framework used by many to bound estimates of the return to education (Manski and Pepper 2000; Okumura and Usui 2014; Mariotti and Meinecke 2015). For instance, when employing a monotone instrumental variable method, they find that the lowest upper bound of the return to schooling is 15.9% for 13–14 years of education and 16.5% for 15–16 years of education.

The partial identification literature also addresses concerns with invalid instruments. For instance, Flores and Flores-Lagunes (2013) derive nonparametric bounds for the local average treatment effect (LATE) without imposing the exclusion restriction assumption. Slichter (2015) bounds estimates of the returns to college using Card's (1995) data. His lower bound is based on the returns of those individuals whose college attendance is unaffected by living close four-year colleges (always takers). His upper bound is computed based on those individuals whose college attendance depends on distance (compliers). Slichter's bounded estimates are between 6.9 and 18.9%.

A significant body of research also examined the impact of school quality on earnings. Card and Krueger (1992) find that higher school quality measured by a lower student–teacher ratio, a longer average term length, and higher teacher pay yield significantly larger returns to schooling for people born between 1920 and 1949. However, in a later paper, Heckman et al. (1996) find that the relationship between school quality and earnings is weak and sensitive to the specification used. Thus, results regarding the impact of school quality are not robust and also are prone to specification biases.

The partial identification bounds estimation approach is also implemented for policy evaluation. For instance, Flores and Flores-Lagunes (2013) and Blanco et al. (2013) estimate bounds for the effect of GED, high school vocational degree, and Job Corps program on earnings. Lee (2009) examines the effect of the Job Corps program on earnings in the presence of sample selection. All these findings suggest that these programs raise earnings for those who participated. Flores and Flores-Lagunes (2013) get a schooling rate of return upper bound of 28% for Job Corps participants.

Partial identification and bounded estimates are nevertheless fallible. They are primarily used to identify causal effects, but can get erroneous parameter estimates if the underlying model is nonlinear. In the human capital model, schooling is nonlinearly related to earnings. A linearized version necessarily omits higher-order schooling terms which are no doubt contained in the error. This linearization is a classic misspecification. As a result, even otherwise valid IVs of schooling yield biased and inconsistent estimates.

9 Early Childhood Development

Our work finds ability to be an important, if not the most important, determinant of earnings. If ability is innate and cannot be changed, then altering the earnings distribution would be impossible. On the other hand, if one can find an intervention to alter ability, then the earnings distribution can be transformed perhaps making it more equal. As Heckman (2008) indicates, one such intervention is investment in early childhood development, and as Tracey and Polachek (2018) show, this result holds even for extremely young children within their first year of life. Cognitive developmental skills, in turn, could boost future earnings. For example, Boissiere et al. (1985), Murnane et al. (1995), Cawley et al. (2001) have demonstrated a positive relationship between cognitive abilities and earnings. Research also shows that a substantial portion of earnings inequality is explained by cognitive abilities (Blau and Kahn 2005).

Studies that focus on non-cognitive abilities also arrive at the same conclusion. Goldsmith et al. (1997) show that self-esteem and locus of control positively influence wages. Kuhn and Weinberger (2004) show that leadership skills positively influence earnings. Muller and Plug (2006) show that the big-five (agreeableness, conscientiousness, extraversion, openness, neuroticism) personality traits influence earnings, with agreeableness having the strongest effect. Finally, Muller and Plug's (2006) paper also finds non-cognitive abilities are as important as cognitive abilities in determining earnings.

Because cognitive and non-cognitive abilities influence the level and distribution of earnings, these type of abilities are important for policy consideration. Some studies argue schooling enhances cognitive skills (Hansen et al. 2004), but a number of other studies emphasize the role of the family. For example, in an early and controversial study, Coleman and his colleagues (1966) highlighted the importance of social capital, namely attributes inherent in the community and family that are useful to the social development of children. Improving resources in the home might be one such initiative. Of course, the other extreme is Herrnstein and Murray (1994) who imply few, if any, interventional benefits.

Recent research links early childhood interventions to boost cognitive- and non-cognitive-type skills. Bowles and Gintis (2002) argue skills can be transferred from previous generations to the next, making the new generation more valuable in the labor market. Based on a randomized experimental setting, Heckman et al. (2006) and Cunha and Heckman (2010) show that family-level intervention during childhood leads to significant improvement

in non-cognitive abilities. A number of other studies (Fletcher and Wolf 2016; Anger and Schnitzlein 2017) also find that family plays an important role in shaping one's cognitive and non-cognitive skills.

Two important issues should be considered to evaluate potential interventions. First is to define the underlying mechanism how family and other factors influence abilities. Second is to assess their economic viability, namely whether the benefits outweigh the associated costs. A number of recent studies address both aspects. Regarding the first, Cunha and Heckman (2007) and Cunha et al. (2010) offer a dynamic structure of skill formation to demonstrate the mechanism through which family and other factors influence children's cognitive and non-cognitive skills. Using Project STAR data on 11,571 kindergarten to third-grade students in Tennessee, Chetty et al. (2011) find small classes increase the likelihood of college attendance many years later. Also, high-scoring classmates lead to higher future earnings, as do more experienced teachers. Further, gains in non-cognitive compared to cognitive skills last longer. Chetty, Friedman, and Rockoff (2014) find that teacher inputs matter. Employing 1989–2009 data on students and teachers in grades 3–8 from a large urban school district, they find the students assigned to a high "value-added" teacher are more likely to attend college, to achieve a higher salary, and less likely to have out of wedlock children. Regarding the second issue, Heckman et al. (2006) and Cunha and Heckman (2010) show that every dollar spent on such childhood interventions yields a 5.7 dollar increase in observed earnings and a projected 8.7 dollar increase in lifetime earnings. These findings reemphasize that appropriate family-level interventions not only enhance abilities and raise earnings, but do so in an economically viable way.

10 Conclusion

Earnings are not uniform across the population. They vary by age, gender, race, and other individual and market characteristics. Many theories evolved to explain earnings. However, in our opinion, the life-cycle human capital approach does best in accounting for the preponderance of these variations. This paper begins by exploring how human capital can explain demographic differences in earnings. In the human capital model, earnings are related to the net stock of human capital an individual accumulates over his or her lifetime. At young ages, when one just enters the labor market and accumulates little human capital, wages are relatively low. At that point, men and women earn comparable wages, but not blacks and whites, most likely because of

school quality differences. Over the life-cycle earnings rise at a diminishing rate, with men's earnings growing more quickly than women's, most likely because of expected differences in lifetime work patterns.

Theory yields a complex nonlinear specification of the earnings function. In the past, this function was too complicated for most researchers to estimate, and still is for intermittent workers. However, the structural model's beauty is its parameterization of previously unmeasured human attributes, specifically three ability measures (two constituting the ability to learn and one constituting the ability to earn), a skill depreciation rate, and a rate of time preference. Unlike IQ and achievement test scores, which have been criticized because they merely assess potential academic accomplishments, these parameters reflect the ability to achieve real-world economic success. Because this structural model directly yields parameters defining rates of time preference, it thereby eliminates the need to perform experimental studies that rely on hypothetical rather than a real-world situation. However, given this model's complex nature, the lack of long enough panel data, algorithmic inefficiencies, and slow computers, virtually all earnings functions emanating from this model have only been estimated population-wide in the aggregate, thus precluding individual-specific values. Nowadays, with new computational technologies and long enough panel data, such functions have finally been estimated person by person.

Our paper makes use of these estimates which vary significantly across the population. A few interesting results emerge when we compare these ability measures with standard IQ values. Whereas these ability measures correlate with IQ-type scores, the correlation between the two is not perfect. Also, the variance of these ability measures is much smaller than the variance in standardized tests. Most of all, racial differences are not as wide. Further, the ability to learn measures are positively related to years of schooling, but the ability to earn is not. Finally, we assess the importance of these new ability measures in explaining earnings variation.

Past analyses estimate a log-linear simplification. This specification, known as the Mincer earnings function, became the workhorse in empirical analysis of earnings determination. Estimated population-wide, and not individual-by-individual, this line of research emphasized schooling as a main determinant of earnings. As a result, numerous studies concentrate on education as a causal impetus. Although these studies show a positive relationship between schooling and earnings, the magnitudes of the estimates differ significantly. Initial OLS analyses yield rates of return that typically range between 5 and 15%, but these estimates are often criticized because schooling is not exogenous, in part because of latent

factors such as unobserved ability. Newer studies rely on instrumental variable techniques to alleviate these biases. However, as Card (2001) reports, the estimates obtained from instrumental variable methods range from 3.6 to 94%.

Such a staggeringly wide range of estimates is not helpful for policymakers. Even if one recognizes that studies examining schooling and earnings use datasets from different countries, years and age cohorts, and rely on different instrumental variables, it is unlikely that the differences in data alone explain such a large variation in the estimates. Rather, it is plausible that the instrumental variables chosen for the estimation may not be fully valid. Many studies show that the IVs used to identify returns to schooling often violate the exclusion restriction, the relevance condition, or both. Of course, the various violations of the assorted IVs can lead to diverse estimates. To unravel these discrepancies, one must understand the underlying structural mechanisms by which the exogenous variations influence the human capital investment process.

Human capital theory postulates that earnings power is determined by accumulated human capital. Schooling emerges as an optimal outcome determined by the relative marginal cost and benefits. The IV-based studies typically identify exogenous variation that influences this decision. But it is perfectly possible that the IVs used, intended solely to measure variation in school, actually influence other aspects of the investment process, as well. The following example illustrates this point. Consider two interventions that cause exogenous variations in years of school: (a) tax credit financial support for education and (b) skill enhancements such as the Perry Preschool Project or Job Corps interventions leading to more education. Each of these interventions can independently serve as an instrument for years of school. Tax credits lower the cost of school attendance, whereas improvements in skill lower the cost of learning leading to more investment in human capital. From a statistical point of view, both would be valid instruments if the interventions are exogenous. As such, they should be able to identify and consistently estimate the causal impact (LATE) of schooling on earnings. However, these interventions can have other implications for investments in human capital. A tax credit helps lower the cost of enrollment and hence only increases the amount of school one obtains for some individuals, and nothing else. On the other hand, an improvement in skills lowers learning costs, thereby increasing years of school, but may also affect post-school investment via the job one gets. In short, the latter instrument may affect a different set of individuals and generate different effects.

Instrumental variables may also generate erroneous estimates for another reason. The human capital model yields a nonlinear earnings–schooling

relationship. Instrumenting the schooling variable in a linear earnings function framework necessarily omits higher-order schooling terms. This omission is a classic misspecification that results in biased and inconsistent estimates. In such a framework, it is impossible to generate a consistent estimate of the returns to schooling even with an instrument that is uncorrelated with other omitted determinants of earnings. It is therefore not possible to fully assess the impact of schooling on earnings without considering the formal structure.

There are efforts (partial identification) to address the potential invalidity of IVs. But most of these efforts make modifications and refinements either based on a given linear functional form or based on nonparametric methods. However, the underlying structural mechanisms still are missing from these analyses. Arguably, these new methodological developments can provide some sense of the estimates by bounding them. But in the absence of an explicit theoretical structure, one cannot be sure the assumptions for bounds (e.g., monotonicity) are necessarily valid.

Another structural aspect that is largely ignored in current empirical work is interpersonal heterogeneity. Heterogeneity essentially means that the functional relationship between the schooling and earnings varies person by person. Estimation without recognizing these structural differences can lead to incomplete and in some cases misleading results. As our preliminary findings show, the results based on the structure and heterogeneity adjusted framework substantially differ from the existing method that does not rely on explicit structures. Contrary to many existing studies, our tentative findings suggest that formal years of schooling plays only a limited role in explaining earnings. In contrast, ability is far more influential in explaining earnings variations. Specifically, one's ability to learn and ability to retain knowledge play the most important roles. This, however, by no means suggests that formal schooling is unimportant. It rather suggests that what is actually learned in school depends on these abilities, so that learning is heterogeneous. Schools may implement ability-enhancing measures which play a role in improving learning outcomes, but merely going to school is not sufficient to learn marketable skills. Indeed, Henderson et al. (2011) find a significant number of individuals with a negative return to schooling. Thus, measures that improve these abilities would be a natural policy intervention to increase earnings and lower earnings disparity.

Acknowledgements The authors thank Thijs ten Raa as well as an anonymous referee for valuable comments that substantially improved the quality of this chapter.

Appendix 1

Optimally producing human capital to maximize lifetime earnings entails equating the marginal costs and marginal benefits of human capital creation in each year of one's life. This process yields a nonlinear (in the parameters) earnings function (Polachek et al. 2015)

$$
\begin{aligned}
Y_t = W^{\frac{1}{(1-b)}} \Bigg[& \left\{ \left(\tfrac{1}{\delta} + \left(E^{1-b} - \tfrac{1}{\delta} \right) e^{\delta(b-1)t^*} \right)^{\frac{1}{(1-b)}} - \left(\tfrac{1}{\delta} \left[\tfrac{b}{r+\delta} \right]^{\frac{b}{(1-b)}} \right) e^{\delta(t^*-t)} \right\} \\
& + \left\{ \frac{1}{\delta \left[\frac{b}{r+\delta} \right]^{\frac{b}{(1-b)}} \left(1 - \frac{b\delta}{r+\delta} \right)} \right\} + \left\{ \left[\tfrac{b}{r+\delta} \right]^{\frac{1}{(1-b)}} \tfrac{1}{(1-b)} e^{(r+\delta)(t-N)} \right\} \\
& - \left\{ 0.5 \left[\tfrac{b}{r+\delta} \right]^{\frac{1}{(1-b)}} \left(\tfrac{1}{(1-b)} \right) \left(\tfrac{b}{(1-b)} \right) e^{2(r+\delta)(t-N)} \right\} \Bigg] + \epsilon_t
\end{aligned}
\tag{9}
$$

where $W = \beta R^{1-b}$, $E = \frac{E_0}{\beta^{\left(\frac{1}{1-b}\right)}}$, t^* is the age at which the individual graduates from school, N is the anticipated retirement age which PDT take to be 65, and E_0 is the human capital stock when training begins. To identify each parameter $b_i, \beta_i, E_{0_i}, r_i, \delta_i$ and R, PDT adopt a two-step process. First, they estimate (9) separately for approximately 1700 individuals in the NLSY79 to obtain b, W, E, d, and r for each individual. Their dependent variable is each of 1700 individual's weekly earnings adjusted to 1982–1984 dollars. The independent variable is each individual's age (t). Years of completed school for each individual are denoted as t^* and remain constant throughout each person's life because PDT's sample omits intermittently schooled individuals.

Second, to identify β_i, E_{0_i}, and the population-wide R, PDT first specify β_i to equal βe_i, where β is the population average and e_i is the individual deviation. Second, they rewrite the $W = \beta R^{1-b}$ coefficient in (9) for individual i as $W_i = R^{1-b_i} \beta e_i$. They then take the logarithm which yields $\ln W_i = (1 - b_i) \ln R + \ln \beta + \ln e_i$ which they then estimate using each individual's values for $\widehat{W}_l$ and $\hat{b}_l$ obtained from estimating (9) above. They obtain β_i by taking the antilog$_e$ of the sum of the latter two terms $\ln \beta + \ln e_i$ in the above equation. Utilizing b_i and β_i values along with the coefficient $\widehat{E}_l = E_{0_i} \Big/ \beta_i^{1/1-b_i}$ obtained from estimating (9) yields individual-specific E_{0_i}. The population-wide rental rate is the coefficient of $(1 - b_i)$.

References

Acemoglu, Daron, and Jorn-Steffen Pischke. 1999. The structure of wages and investment in general training. *Journal of Political Economy* 107 (3): 539–572.

Albrecht, J.W., P.A. Edin, M. Sundstrom, and S.B. Vroman. 1999. Career interruptions and subsequent earnings: A reexamination using Swedish data. *Journal of Human Resources* 34 (2): 294–311.

Anger, Silke, and Daniel Schnitzlein. 2017. Cognitive skills, non-cognitive skills, and family background: Evidence from sibling correlations. *Journal of Population Economics* 30 (2): 591–620.

Angrist, Joshua, and Alan Krueger. 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106 (4): 979–1014.

Armenter, Roc. 2015. A bit of a miracle no more: The decline of the labor share. *Business Review*, Third Quarter, Federal Reserve Bank of Philadelphia Research Department.

Ashenfelter, Orley, and Alan Krueger. 1994. Estimating the returns to schooling using a new sample of twins. *American Economic Review* 84: 1157–1173.

Ashenfelter, Orley, and Cecilia Rouse. 1998. Income, schooling, and ability: Evidence from a new sample of identical twins. *The Quarterly Journal of Economics* 113 (1): 253–284.

Ashenfelter, Orley, C. Harmon, and H. Oosterbeek (eds.). 1999. Economic returns to schooling: New evidence. Special Issue of *Labour Economics* 6 (4).

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John van Reenen. 2017. Concentrating on the fall of the labor share, IZA DP #10539.

Baldwin, M.L., L.A. Zeager, and P.R. Flacco. 1994. Gender differences in wage losses from impairments: Estimates from the survey of income and program participation. *Journal of Human Resources* 29 (3): 865–887.

Barro, Robert, and X. Sala-i-Martin. 1999. *Economic growth*. Cambridge, MA: MIT Press.

Baum, Charles L. 2002. The effect of work interruptions on women's wages. *Labour* 16 (1): 1–36.

Becker, Gary. 1964. *Human capital: A theoretical and empirical analysis, with special preferences to education*. Chicago: University of Chicago Press.

Becker, Gary, and Barry Chiswick. 1966. Education and the distribution of earnings. *American Economic Review* 56: 358–369.

Benmelech, Efraim, and Claude Berrebi. 2006. Attack assignments in terror organizations and the productivity of suicide bombers. Working Paper, Harvard University.

Ben-Porath, Yoram. 1967. The production of human capital over the life cycle. *Journal of Political Economy* 75: 352–365.

Bergmann, Barbara R. 1971. The effect on white incomes of discrimination in employment. *Journal of Political Economy* 79 (2): 294–313.

Bergmann, Barbara R. 1974. Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal* 1 (2): 103–110.

Bhuller, Manudeep, Magne Mogstad, and Kjell Salvanes. 2014. Life cycle earnings, education premiums and internal rates of return. NBER Working Papers no. 20250.

Bishop, John. 1997. What we know about employer provided training: A review of the literature. *Research in Labor Economics* 16: 19–88.

Blanco, German, Carlos A. Flores, and Alfonso Flores-Lagunes. 2013. Bounds on average and quantile treatment effects of job corps training on wages. *Journal of Human Resources* 48 (3): 659–701.

Blau, Francine, and Lawrence M. Kahn. 1992. The gender earnings gap: Learning from international comparisons. *American Economic Review* 82 (2): 533–538.

Blau, Francine, and Lawrence M. Kahn. 2005. Do cognitive test scores explain higher U.S. wage inequality? *The Review of Economics and Statistics* 87 (1): 184–193.

Blundell, Richard. 2014. Income dynamics and life-cycle inequality: Mechanisms and controversies. *The Economic Journal* 124: 289–318.

Boissiere, M., J.B. Knight, and R.H. Sabot. 1985. Earnings, schooling, ability, and cognitive skills. *The American Economic Review* 75 (5): 1016–1030.

Booth, Alison L., and Pamela Katic. 2013. Cognitive skills, gender and risk preferences. *Economic Record* 89 (284): 19–30.

Borjas, George J. 1982. The earnings of male Hispanic immigrants in the United States. *Industrial and Labor Relations Review* 35 (3): 343–353.

Borjas, George J. 1985. Assimilation, changes in cohort quality, and the earnings of immigrants. *Journal of Labor Economics* 3 (4): 463–489.

Borjas, George J. 1993. The intergenerational mobility of immigrants. *Journal of Labor Economics* 11 (1): 113–135.

Bound, John, and David A. Jaeger. 1996. On the validity of season of birth as an instrument in wage equations: A comment on Angrist & Krueger's "Does compulsory school attendance affect schooling and earnings". NBER Working Paper 5835.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak. *Journal of the American Statistical Association* 90 (430): 443–450.

Bowles, Samuel, and Herbert Gintis. 2002. The inheritance of inequality. *Journal of Economic Perspectives* 16 (3): 3–30.

Bowles, Samuel, Herbert Gintis, and Melissa Osborne. 2001. The determinants of earnings: A behavioral approach. *Journal of Economic Literature* 39 (4): 1137–1176.

Bowlus, Audra, and Chris Robinson. 2012. Human capital Prices, productivity, and growth. *American Economic Review* 102 (7): 3483–3515.

Brown, Charles. 1980. The 'overtaking' point revisited. *Review of Economics and Statistics* 62 (2): 309–313.

Brown, Heather, and Marjon van der Pol. 2015. Intergenerational transfer of time and risk preferences. *Journal of Economic Psychology* 49: 187–204.

Browning, Martin, and Mette Ejrnæs. 2013. Heterogeneity in the dynamics of labor earnings. *Annual Review of Economics* 5: 219–245.

Browning, Martin, Lars Hansen, and James Heckman. 1999. Micro data and general Equilibrium models. In *Handbook of macroeconomics*, vol. 1A, ed. John Taylor and Michael Woodford, 543–633. Amsterdam: Elsevier.

Browning, Martin, Mette Ejrnæs, and Javier Alvarez. 2010. Modelling income processes with lots of heterogeneity. *Review of Economic Studies* 77 (4): 1353–1381.

Burgess, Simon, and Carol Propper. 1998. Early health-related behaviours and their impact on later life chances: Evidence from the US. *Health Economics* 7 (5): 381–399.

Card, David. 1995. Using geographic variation in college proximity to estimate the return to schooling. In *Aspects of labor market behaviour: Essays in honour of John Vanderkamp*, ed. L.N. Christofides, E.K. Grant, and R. Swidinsky. Toronto: University of Toronto Press.

Card, David. 1999. The causal effect of education on earnings. In *Handbook of labor economics*, vol. 3A, ed. O. Ashenfelter and D. Card. Amsterdam: Elsevier.

Card, David. 2001. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica* 69 (5): 1127–1160.

Card, David, and Alan B. Krueger. 1992. School quality and black-white relative earnings: A direct assessment. *Quarterly Journal of Economics* 107 (1): 151–200.

Carneiro, Pedro, and James Heckman. 2002. The evidence on credit constraints in post-secondary schooling. *The Economic Journal* 112: 989–1018.

Catalyst. 2003. Workplace flexibility is still a women's advancement issue. http://64.233.167.104/u/Catalyst?q=cache:BGumQKH8saEJ:www.catalystwomen.org/bookstore/files/view/Workplace%2520Flexibility%2520Is%2520Still%2520a%2520Women%27s%2520Advancement%2520Issue.pdf+mba+and+men+and+women&hl=en&ie=UTF-8.

Cawley, John, James Heckman, and Edward Vytlacil. 2001. Three observations on wages and measured cognitive ability. *Labour Economics* 8 (4): 419–442.

Chetty, Raj, John N. Friedman, and Jonah Rockoff. 2014. Measuring the impact of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review* 104 (9): 2633–2679.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. How does your kindergarten classroom affect your earnings? Evidence from project star. *Quarterly Journal of Economics* 126 (4): 1593–1660.

Chiswick, Barry R. 1978. The effect of americanization on the earnings of foreign-born men. *Journal of Political Economy* 86 (5): 897–921.

Chiswick, Barry, and Jacob Mincer. 1972. Time-series changes in personal income inequality in the United States. *Journal of Political Economy* 80 (3): S34–S66.

Clark, Damon, and Paco Martorell. 2014. The signaling value of a high school diploma. *Journal of Political Economy* 122 (2): 282–318.

Cobb-Clark, Deborah A., and Michelle Tan. 2011. Noncognitive skills, occupational attainment, and relative wages. *Labour Economics* 18 (1): 1–13.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. 1966. *Equality of education opportunity*. U.S. Department of Health, Education, and Welfare, Office of Education, U.S. Government Printing Office Washington.

Conneely, K., and R. Uusitalo. 1997. Estimating heterogeneous treatment effects in the Becker schooling model. Industrial Relations Section, Princeton University.

Corcoran, Mary, and Greg Duncan. 1979. Work history, labor force attachment, and earnings differences between the races and sexes. *Journal of Human Resources* 14 (1): 3–20.

Corcoran, Mary, Greg Duncan, and Michael Ponza. 1983. Longitudinal analysis of white women's wages. *Journal of Human Resources* 18 (4): 497–520.

Cortès, Patricia, and Jessica Pan. 2016. Prevalence of long hours and women's job choices: Evidence across countries and within the U.S. Paper presented at the 2016 ASSA Annual Conference.

Croson, Rachel, and Uri Gneezy. 2009. Gender differences in preferences. *Journal of Economic Literature* 47 (2): 448–474.

Cseh, Attila. 2008. The effects of depressive symptoms on earnings. *Southern Economic Journal* 75 (2): 383–409.

Cunha, Flavio, and James Heckman. 2007. The technology of skill formation. *The American Economic Review* 97 (2): 31–47.

Cunha, Flavio, and James Heckman. 2010. Investing in our young people. NBER Working Paper No. 16201.

Cunha, Flavio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov. 2006. Interpreting the evidence on life cycle skill formation. In *Handbook of the economics of education*, ed. Eric Hanushek and Finis Welch, 698–812. Amsterdam: North Holland.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica* 78 (3): 883–931.

Czarnitzki, Dirk, and Thorsten Doherr. 2009. Genetic algorithms for econometric optimization. Working Paper.

Dechter, Evgenia K. 2015. Physical appearance and earnings, hair color matters. *Labour Economics* 32: 15–26.

Dohmen, Thomas, and Armin Falk. 2001. Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review* 101 (2): 556–590.

Dorsey, Robert, and Walter Mayer. 1995. Genetic algorithms for estimation problems with multiple optima, nondifferentiability, and other irregular features. *Journal of Business & Economic Statistics* 13 (1): 53–66.

Duflo, Esther. 2001. Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment. *American Economic Review* 91 (4): 795–813.

Elsby, Mike, Bart Hobijn, and Aysegul Sahin. 2013. The decline of the U.S. labor share. *Brookings Papers on Economic Activity*, 1–42.

Fletcher, Jason M. 2014. The effects of childhood ADHD on adult labor market outcomes. *Health Economics* 23 (2): 159–181.

Fletcher, Jason M., and Barbara Wolfe. 2016. The importance of family income in the formation and evolution of non-cognitive skills in childhood. La Follette School Working Paper no. 2016-001, University of Wisconsin.

Flores, C., and Alfonso Flores-Lagunes. 2013. Partial identification of local average treatment effects with an invalid instrument. *Journal of Business and Economic Statistics* 31 (4): 534–545.

Fryer, R. G., Jr., and S. D. Levitt. 2004. Understanding the black-white test score gap in the first two years of school. *The Review of Economics and Statistics* 86 (2): 447–464.

Fryer, Roland, G. Jr., and Steven D. Levitt. 2013. Testing for racial differences in the mental ability of young children. *American Economic Review* 103 (2): 981–1005.

Fuchs, Victor. 1967. Hourly earnings differentials by region and size of city. *Monthly Labor Review* 94 (5): 9–15.

Gabriel, Paul E. 2016. The doughboy premium: An empirical assessment of the relative wages of American veterans of World War I. *Applied Economics Letters* 23 (2): 93–96.

Gibson, John, and Osaiasi Koliniusi Fatai. 2006. Subsidies, selectivity and the returns to education in urban Papua New Guinea. *Economics of Education Review* 25 (2): 133–146.

Goldberg, David. 1989. *Genetic algorithms in search, optimization, and machine learning*. Reading, MA: Addison-Wesley.

Goldin, Claudia. 2014. A grand gender convergence: Its last chapter. *American Economic Review* 104 (4): 1091–1119.

Goldin, Claudia, and Solomon Polachek. 1987. Residual differences by sex: Perspectives on the gender gap in earnings. *American Economic Review* 77 (2): 143–151.

Goldsmith, Arthur, Jonathan Veum, and William Darity. 1997. Unemployment, joblessness, psychological well-being and self-esteem: Theory and evidence. *Journal of Socio-Economics* 26 (2): 133–158.

Greene, William. 2005. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics* 126 (2): 269–303.

Greene, William. 2010. Distinguishing between heterogeneity and inefficiency: Stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Economics* 13 (10): 959–980.

Gronau, Reuben. 1988. Sex-related wage differentials and women's interrupted labor careers-The chicken or the egg. *Journal of Labor Economics* 6 (3): 277–301.

Grossman, Michael. 1972. On the concept of health capital and the demand for health. *The Journal of Political Economy* 80 (2): 223–255.

Groves, Melissa O. 2005. How important is your personality? Labor market returns to personality for women in the US and UK. *Journal of Economic Psychology* 26 (6): 827–841.

Haley, William J. 1976. Estimation of the earnings profile from optimal human capital accumulation. *Econometrica* 44 (6): 1223–1238.

Hamermesh, Daniel, and Jeff E. Biddle. 1994. Beauty and the labor market. *American Economic Review* 84: 1174–1194.

Hamermesh, Daniel, Xin Meng, and Junsen Zhang. 2002. Dress for success—Does primping pay? *Labour Economics* 9 (3): 361–373.

Hammitt, James K., and Tuba Tuncel. 2015. Preferences for life-expectancy gains: Sooner or later? *Journal of Risk and Uncertainty* 51 (1): 79–101.

Hansen, Karsten T., James Heckman, and Kathleen J. Mullen. 2004. The effect of schooling and ability on achievement test scores. *Journal of Econometrics* 121 (1–2): 39–98.

Harmon, Colm, and Ian Walker. 1995. Estimates of the economic return to schooling for the United Kingdom. *American Economic Review* 85 (5): 1278–1286.

Hartog, Joop, Gerard Pfann, and Geert Ridder. 1989. (Non-)graduation and the earnings function: An inquiry on self-selection. *European Economic Review* 33 (7): 1373–1395.

Heckman, James. 1975. Estimates of aging capital production function embedded in a life-cycle model of labor supply. In *Household production and consumption*, ed. Nester Terleckyj, 227–258. New York: Columbia University Press for the National Bureau of Economic Research.

Heckman, James. 1976. A life-cycle model of earnings, learning, and consumption. *Journal of Political Economy* 84 (4): S11–S44.

Heckman, James. 2008. School, skills, and synapses. *Economic Inquiry* 46 (3): 289–324.

Heckman, James, Anne Layne-Farrar, and Petra Todd. 1996. Human capital pricing equations with an application to estimating the effect of schooling quality on earnings. *The Review of Economics and Statistics* 78 (4): 562–610.

Heckman, James, Lance Lochner, and Christopher Taber. 1998. Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. *Review of Economic Dynamics* 1: 1–58.

Heckman, James, Lance Lochner, and Petra Todd. 2003. Fifty years of Mincer earnings regressions. IZA Discussion Paper No. 775.

Heckman, James, Lance Lochner, and Petra Todd. 2006. Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. In *Handbook of the economics of education*, ed. Eric A. Hanushek and Finis Welch, 307–458. Amsterdam: North Holland.

Heckman, James, and Solomon Polachek. 1974. The functional form of the income-schooling relation. *Journal of the American Statistical Association* 69: 350–354.

Henderson, Daniel, Solomon Polachek, and Le Wang. 2011. Heterogeneity in schooling rates of return. *Economics of Education Review* 30 (6): 1202–1214.

Herrnstein, R. J., and C. Murray. 1994. *The bell curve: Intelligence and class structure in American life*. New York: The Free Press.

Hoffmann, Florian. 2016. HIP, RIP and the robustness of empirical earnings processes. Vancouver School of Economics, University of British Columbia Version.

Holland, J.H. 1975. *Adaptation in natural and artificial systems*. Ann Arbor, MI: University of Michigan Press.

Hosios, Arthur. 1990. On the efficiency of matching and related models of search and unemployment. *The Review of Economic Studies* 57 (2): 279–298.

Hotchkiss, Julie, and Melinda Pitts. 2003. At what level of labour market intermittency are women penalized? *American Economic Review Papers and Proceedings* 93 (2): 233–237.

Hotchkiss, Julie, and Melinda Pitts. 2005. Female labour force intermittency and current earnings: Switching regression model with unknown sample selection. *Applied Economics* 37 (5): 545–560.

Hull, Charles R. (ed.). 1899. *The economic writings of Sir William Petty*. Cambridge: Cambridge University Press.

Ichino, Andrea, and Rudolf Winter-Ebmer. 2004. The long-run educational cost of World War II. *Journal of Labor Economics* 22 (1): 57–86.

Imbens, Guido, and Paul Rosenbaum. 2005. Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society A* 168 (1): 109–126.

Jacobsen, J.P., and L.M. Levin. 1995. Effects of intermittent labour force attachment on women's earnings. *Monthly Labour Review*, September 14–19.

Johnson, Thomas. 1970. Returns from investment in human capital. *American Economic Review* 60 (4): 546–560.

Johnson, Thomas. 1978. Time in school: The case of the prudent patron. *American Economic Review* 68 (5): 862–872.

Johnson, Thomas, and Frederick Hebein. 1974. Investments in human capital and growth in personal income 1956–1966. *American Economic Review* 64 (4): 604–615.

Jovanovic, Boyan. 1979. Job matching and the theory of turnover. *Journal of Political Economy* 87 (5): 972–990.

Kane, Thomas J., and Cecilia E. Rouse. 1995. Labor market returns to two- and four-year colleges: Is a credit a credit and do degrees matter? *American Economic Review* 85 (3): 600–614.

Kao, Charng, Solomon Polachek, and Phanindra Wunnava. 1994. Male-female wage differentials in Taiwan: A human capital approach. *Economic Development and Cultural Change* 42 (2): 351–374.

Karabarbounis, Loukas, and Brent Neiman. 2013. The global decline of the labor share. *Quarterly Journal of Economics* 129 (1): 61–103.
Kiker, B.F. 1966. The historical roots and the concept of human capital. *Journal of Political Economy* 74: 481–799.
Kiker, B.F., and M. Mendes de Oliveira. 1992. Optimal allocation of time and estimation of market wage functions. *Journal of Human Resources* 27 (3): 445–471.
Kim, Moon K., and Solomon Polachek. 1994. Panel estimates of male-female earnings functions. *Journal of Human Resources* 27 (3): 445–471.
King, Robert, and Sergio Rebelo. 1999. Resuscitating real business cycles. In *Handbook of macroeconomics*, vol. 1B, ed. John Taylor and Michael Woodford, 927–1007. Amsterdam: Elsevier.
Kitagawa, Toru. 2015. A test for instrument validity. *Econometrica* 83 (5): 2043–2063.
Klawitter, Marieka. 2015. Meta-analysis of the effects of sexual orientation on earnings. *Industrial Relations* 54 (1): 4–32.
Kleibergen, F. 2002. Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* 70: 1781–1803.
Korenman, Sanders, and David Neumark. 1992. Marriage, motherhood, and wages. *Journal of Human Resources* 27 (2): 233–255.
Krueger, Alan. 1999. Measuring labor's share. *American Economic Review, Papers and Proceedings* 89 (2): 45–51.
Kuhn, Peter, and Catherine Weinberger. 2004. Leadership skills and wages. Santa Barbara Working Paper, University of California. http://econ.ucsb.edu/~weinberg/Leader.pdf.
Kumbhakar, Subal. 1996. A farm-level study of labor use and efficiency wages in Indian agriculture. *Journal of Econometrics* 72: 177–195.
Kuratani, Masatoshi. 1973. A *theory of training, earnings and employment in japan*. Ph.D. dissertation, Columbia University.
Lazear, Edward. 1995. *Personnel economics*. Cambridge, MA: The MIT Press.
Lazear, Edward, and Sherwin Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89 (5): 841–864.
Lazear, Edward, and Sherwin Rosen. 1990. Male-female wage differentials in job ladders. *Journal of Labor Economics* 8 (1): S106–S123.
Lee, David S. 2009. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies* 76 (3): 1071–1102.
Lemieux, Thomas, and David Card. 1998. Education, earnings, and the 'Canadian G.I. Bill'. National Bureau of Economic Research Working Paper no. 6718.
Lewis, H. Gregg. 1963. *Unionism and relative wages in the United States: An empirical inquiry*. Chicago: University of Chicago Press.
Lewis. H. Gregg. 1986. *Union relative wage effects: A survey*. Chicago: University of Chicago Press.
Licht, G., and V. Steiner. 1991. Male-female wage differentials, labor force attachment, and human capital accumulation in Germany. Working Paper no. 65 (institut fr Volkswirtschafslehre der Universit Augburf).

Light, Audrey, and M. Ureta. 1995. Early-career work experience and gender wage differentials. *Journal of Labor Economics* 13 (1): 121–154.

Liu, Huju. 2009. Life cycle human capital formation, search intensity and wage dynamics. Working Paper, University of Western Ontario.

Lochner, Lance, and Enrico Moretti. 2004. The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review* 94 (1): 155–189.

Lundborg, Petter, Paul Nystedt, and Dan-Olof Rooth. 2014. Height and earnings: The role of cognitive and noncognitive skills. *Journal of Human Resources* 49 (1): 141–166.

Magnac, Thierry, Nicolas Pistolesi, and Sébastien Roux. 2018. Post-schooling human capital investments and the life cycle of earnings. *Journal of Political Economy* 126 (3): 1219–1249.

Maluccio, John. 1997. Endogeneity of schooling in the wage Function. Working Paper, Yale University Department of Economics.

Manski, Charles F., and John V. Pepper. 2000. Monotone instrumental variables: With an application to the returns to schooling. *Econometrica* 689 (4): 997–1010.

Manski, Charles F., and John V. Pepper. 2009. More on monotone instrumental variables. *The Econometrics Journal* 12 (1): S200–S216.

Mariotti, Martine, and Juergen Meinecke. 2015. Partial identification and bound estimation of the average treatment effect of education on earnings for South Africa. *Oxford Bulletin of Economics and Statistics* 77 (2): 210–233.

Meghir, Costas, and Luigi Pistaferri. 2011. Earnings, consumption and life cycle choices. In *Handbook of labor economics*, vol. 4B, ed. O. Ashenfelter and D. Card, 774–854. Amsterdam: Elsevier North Holland.

Meghir, Costas, and Mårten Palme. 1999. Assessing the effect of schooling on earnings using a social experiment. IFS Working Papers no. W99/10, Institute for Fiscal Studies.

Michael, Robert. 1973. Education in nonmarket production. *Journal of Political Economy* 81 (2): 306–327.

Miller, Carloe. 1993. Actual experience, potential experience or age, and labor force participation by married women. *Atlantic Economic Journal* 21 (4): 60–66.

Mincer, Jacob. 1958. Investment in human capital and the personal income distribution. *Journal of Political Economy* 66: 281–302.

Mincer, Jacob. 1974. *Schooling, experience, and earnings*. New York: Columbia University Press for the National Bureau of Economic Research.

Mincer, Jacob, and Haim Ofek. 1982. Interrupted work careers: Depreciation and restoration of human capital. *Journal of Human Resources* 17: 3–24.

Mincer, Jacob, and Solomon Polachek. 1974. Family investments in human capital. *Journal of Political Economy* 82: S76–S108.

Mincer, Jacob, and Solomon Polachek. 1978. Women's earnings reexamined. *Journal of Human Resources* 13 (1): 118–134.

Montenegro, Claudio E., and Harry Anthony Patrinos. 2014. Comparable estimates of returns to schooling around the world. Policy Research Working Paper no. WPS 7020, World Bank Group, Washington, DC.

Moreira, Marcelo J. 2003. A conditional likelihood ratio test for structural models. *Econometrica* 71 (4): 1027–1048.

Mueller, Gerrit, and Erik J.S. Plug. 2006. Estimating the effect of personality on male and female earnings. *Industrial and Labor Relations Review* 60 (1): 3–22.

Murnane, R.J., J.B. Willett, and F. Levy. 1995. The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics* 77 (2): 251–266.

Murphy, Kevin, and Finis Welch. 1990. Empirical age-earnings profiles. *Journal of Labor Economics* 8: 202–229.

Okumura, Tsunao, and Emiko Usui. 2014. Concave-monotone treatment response and monotone treatment selection: With an application to the returns to schooling. *Quantitative Economics* 5 (1): 175–194.

Oreopoulos, Philip, and Uros Petronijevic. 2013. Making college worth it: A review of research on the returns to higher education. NBER Working Paper no. 19053.

Patrinos, H. A., and G. Psacharopoulos. 2010. Returns to education in developing countries. In *International encyclopedia of education*, ed. P. Penelope, B. Eva, and M. Barry, 305–312. Oxford: Elsevier.

Phipps, S., P. Burton, and L. Lethbridge. 2001. In and out of the labour market: Long-term income consequences of child-related interruptions to women's paid work. *Canadian Journal of Economics* 34 (2): 411–429.

Polachek, Dora, and Solomon Polachek. 1989. An indirect test of children's influence on efficiencies in parental consumer behavior. *Journal of Consumer Affairs* 23 (1): 91–110.

Polachek, Solomon. 1975a. Differences in expected post-school investment as a determinant of market wage differentials. *International Economic Review* 16: 451–470.

Polachek, Solomon. 1975b. Potential biases in measuring discrimination. *Journal of Human Resources* 6: 205–229.

Polachek, Solomon. 1979. Occupational segregation among women: Theory, evidence, and a prognosis. In *Women in the labor market*, ed. C. Lloyd, E. Andrews, and C. Gilroy, 137–157. New York: Columbia University Press.

Polachek, Solomon. 1981. Occupational self-selection: A human capital approach to sex differences in occupational structure. *Review of Economics and Statistics* 63 (1): 60–69.

Polachek, Solomon. 1987. Occupational segregation and the gender wage gap. *Population Research and Policy Review* 6: 47–67.

Polachek, Solomon. 2003. Mincer's overtaking point and the life cycle earnings distribution. *Review of Economics of the Household* 1: 273–304.

Polachek, Solomon. 2008. Earnings over the life cycle: The Mincer earnings function and Its applications. *Foundations and Trends in Microeconomics* 4 (3): 165–272.

Polachek, Solomon. 2012. Introduction to a life cycle approach to migration: Analysis of the perspicacious peregrinator. *Research in Labor Economics* 35: 341–347.

Polachek, Solomon, Tirthatanmoy Das, and Rewat Thamma-Apiroam. 2015. Micro- and macroeconomic implications of heterogeneity in the production of human capital. *Journal of Political Economy* 123 (6): 1410–1455.

Polachek, Solomon, and Francis Horvath. 2012. A life cycle approach to migration: Analysis of the perspicacious peregrinator. *Research in Labor Economics* 35: 349–395.

Polachek, Solomon, and John Robst. 2001. Trends in the male-female wage gap: The 1980s compared with the 1970s. *Southern Economic Journal* 67 (4): 869–888.

Psacharopoulos, George. 2006. The value of investment in education: Theory, evidence, and policy. *Journal of Education Finance* 32 (2): 113–136.

Psacharopoulos, George, and Harry Anthony Patrinos. 2004. Returns to investment in education: A further update. *Education Economics* 12 (2): 111–134.

Rai, Jyoti, and Jean Kimmel. 2015. Gender differences in risk preferences: An empirical study using attitudinal and behavioral specifications of risk aversion. *Research in Labor Economics* 42: 61–92.

Riddell, W. Craig, and Xueda Song. 2017. The role of education in technology use and adoption: Evidence from the Canadian workplace and employee survey. *Industrial and Labor Relations Review* 70 (5): 1219–1253.

Robins, Philip, Jenny F. Homer, and Michael T. French. 2011. Beauty and the labor market: Accounting for the additional effects of personality and grooming. *Labour* 25 (2): 228–251.

Rosen, Sherwin. 1976. A theory of life earnings. *Journal of Political Economy* 84 (4): S45–S67.

Rosen, Sherwin. 1981. The economics of superstars. *American Economic Review* 71 (5): 845–858.

Roy, A.D. 1950. The distribution of earnings and of individual output. *Economic Journal* 60 (239): 489–505.

Rummery, S. 1992. The contribution of intermittent labour force participation to the gender wage differential. *Economic Record* 68 (203): 351–364.

Sabia, Joseph J. 2015. Fluidity in sexual identity, unmeasured heterogeneity, and the earnings effects of sexual orientation. *Industrial Relations* 54 (1): 33–58.

Saiz, Albert, and Elena Zoido. 2005. Listening to what the world says: Bilingualism and earnings in the United States. *The Review of Economics and Statistics* 87 (3): 523–538.

Sandell, Steven, and David Shapiro. 1980. Work expectations, human capital accumulation and the wages of young women. *Journal of Human Resources* 15 (3): 335–353.

Scholz, John Karl, and Kamil Sicinski. 2015. Facial attractiveness and lifetime earnings: Evidence from a cohort study. *The Review of Economics and Statistics* 97 (1): 14–28.

Sen, B. 2001. Revisiting women's preferences about future labor force attachment: What effects do they have on earnings and what are they affected by? *Worker Wellbeing in a Changing Labor Market, Research in Labor Economics* 20: 311–337.

Simpson, Wayne. 2000. Intermittent work activities and earnings. *Applied Economics* 32 (14): 1777–1786.

Slichter, David P. 2015. The employment effects of the minimum wage: A selection ratio approach to measuring treatment effects. Working Paper.

Song, Xueda, and John Jones. 2006. The effects of technological change on life-cycle human capital investment. Working Paper.

Spivey, Christy. 2005. Time off at what price? The effects of career interruptions on earnings. *Industrial and Labor Relations Review* 59 (1): 119–140.

Stafford, Frank, and Marianne Sundstrom. 1996. Time out for childcare: Signalling and earnings rebound effects for men and women. *Labour* 10 (3): 609–629.

Staiger, Douglas, and James H. Stock. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65 (3): 557–586.

Steen, Todd. 2004. The relationship between religion and earnings: Recent evidence from the NLS Youth Cohort. *International Journal of Social Economics* 31 (5/6): 572–581.

Sternberg, R.J. 1985. *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Stratton, Leslie. 1995. The effect of interruptions in work experience have on wages. *Southern Economic Journal* 61 (4): 955–970.

Suter, Larry E., and Herman P. Miller. 1973. Income difference between men and career women. *American Journal of Sociology* 78 (4): 962–974.

Theeuwes, J., C. Koopmans, R. Van Opstal, and H. Van Reijn. 1985. Estimation of optimal human capital accumulation parameters for the Netherlands. *European Economic Review* 29 (2): 233–257.

Tracey, Marlon, and Solomon Polachek. 2018. If looks could heal: Child health and paternal investment. *Journal of Health Economics* 57: 179–190.

Trostel, Philip, Ian Walker, and Paul Woolley. 2002. Estimates of the economic return to schooling for 28 countries. *Labour Economics* 9 (1): 1–16.

Tucker-Drob, Elliot. 2009. Differentiation of cognitive abilities across the life span. *Developmental Psychology* 45 (4): 1097–1118.

Verdon, Andrew. 2018. Human capital: Homogenous or heterogeneous? An empirical test. Working Paper, Binghamton University.

Von Weizsäcker, Robert. 1993. *A theory of earnings distribution*. Cambridge: Cambridge University Press.

Wallace, T. Dudley, and Lauren Ihnen. 1975. Full-time schooling in life cycle models of human capital accumulation. *Journal of Political Economy* 83 (1): 137–155.

Walsh, John. 1935. Capital concept applied to man. *Quarterly Journal of Economics* 49: 255–285.

Webber, Douglas. 2014. The lifetime earnings premia of different majors: Correcting for selection based on cognitive, noncognitive, and unobserved factors. *Labour Economics* 28: 14–23.

Weinberger, Catherine J. 2014. The increasing complementarity between cognitive and social skills. *The Review of Economics and Statistics.* 96 (5): 849–861.

Weiss, Yoram, and Reuben Gronau. 1981. Expected interruptions in labor force participation and sex related differences in earnings growth. *Review of Economic Studies* 48 (4): 607–619.

Welch, Finis. 1974. Black white differences in returns to schooling. *American Economic Review* 63 (5): 893–907.

Wiswall, Matthew, and Basit Zafar. 2016. Preference for the workplace, human capital, and gender. NBER Working Papers no. 22173.

Wu, Houying. 2007. Can the human capital approach explain life-cycle wage differentials between races and sexes? *Economic Inquiry* 45 (1): 24–39.

Performance: The Output/Input Ratio

Thijs ten Raa

1 Introduction

A production unit, organization, firm, industry, or economy performs well if it produces much output per unit of input, in other words, when the output/input ratio is high. The main performance measure is productivity. There are subtle connections between performance, productivity, efficiency, and profitability. Analysis of their interrelations will take us through many issues and concepts of measurement and will connect different bodies of literature, namely in economics and operations research.

The measurement of performance using an output/input ratio presumes that output components and input components can each be aggregated. This is particularly true for inputs. Production requires multiple inputs, typically labor and capital services. On the output side, the aggregation issue is often circumvented. One way is to break down production in micro-activities, one for each type of output. This approach moves the aggregation issue away from commodities toward the micro-performance measures (Blackorby and Russell 1999). An obvious alternative way to circumvent output aggregation is to assume that there is a single performance criterion, such as profit, but this approach raises the question if profit is a better measure of performance than, say, real profit (profit divided by a price index). A windfall profit due to a price shock, without any change in the input-output

T. ten Raa (✉)
Utrecht School of Economics, Utrecht University, Utrecht, The Netherlands
e-mail: tenraa@uvt.nl

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_3

structure of production, does not reflect an improvement in management performance. In other words, we better disentangle profit in a real performance component and a nominal price effect. This issue is related to the design of bonus schedules for managers, where profit is shown to be a good proxy for effort only if the distribution of the windfall component fulfills a certain property (the likelihood ratio monotonicity of Milgrom 1981).

Throughout this chapter, I assume constant returns to scale, unless explicitly stated otherwise. With increasing complexity, I will discuss, first, single input-single output production; second, multiple input-single output production; third, single input-multiple output production; and, fourth, multiple input-multiple output production. The simplest of these cases, single input-single output production, suffices to discuss the relationship between performance, productivity, efficiency, and profitability.

Consider a single input-single output industry with two firms, a duopoly. Denote the input quantities by x and the output quantities by y. Use superscripts to indicate to which firm a symbol pertains: firm 1 or firm 2. Let the first firm be the more productive than the second: $y^1/x^1 > y^2/x^2$. (This is an innocent assumption, because we are free to relabel the firms.) Then firm 1 can produce no more than it produces, at least under the assumptions that the data represent all conceivable practices of production and that the firm's input is fixed. Firm 2, however, could perform better by adopting the production technique of firm 1. That way it would produce y^1/x^1 units per unit of input and since it commands x^2 inputs, its potential output is $(y^1/x^1)\ x^2$. By the presumed productivity inequality, this exceeds the actually produced quantity, y^2.

In our discussion, we must distinguish observed market prices and competitive shadow prices. Market prices are observed and may vary. Some firms negotiate tighter labor conditions than others, and some firms may have shrewder salesmen, extracting higher prices. Someone who "could sell sand to the Arabs" exercises market power but is not productive; the market price exceeds the production price. Production prices are shadow prices which in turn are associated with the constraints of a program that determines the optimum allocation of resources. Later on, optimality will be linked to the consumer's preferences, but in the introductory Mickey Mouse duopoly, it reduces to the maximization of output subject to an input constraint. The maximization program can be applied to a firm (1 or 2) and to the industry (the duopoly), to determine firm and industry efficiencies. The simplest program features constant returns to scale and is applied to a firm, say firm 1:

$$\max_{\theta_1,\theta_2,c\geq 0} y^1 c : x^1\theta_1 + x^2\theta_2 \leq x^1, y^1\theta_1 + y^2\theta_2 \geq y^1 c. \tag{1}$$

In program (1), firm 1 runs activities 1 (input x^1, output y^1) and 2 (input x^2, output y^2) with intensities θ_1 and θ_2, respectively, and c is the expansion factor for output. The first constraint binds the required input by the available input. Denote the Lagrange multiplier or shadow price of this constraint by w (the labor wage). The second constraint binds the expanded output by the sum of the activity outputs. Denote the Lagrange multiplier or shadow price of this constraint by p (the product price). The shadow prices are relevant for performance measurement and are the variables of the dual program associated with the primal program, (1) in this case.

The dual program minimizes the value of the bounds subject to the dual constraint. The bounds are x^1 and 0, so the objective of the dual program is wx^1 or, equivalently, w. The dual constraint is $(w\ p)\begin{pmatrix} x^1 & x^2 & 0 \\ -y^1 & -y^2 & y^1 \end{pmatrix} \geq (0\ 0\ y^1)$, featuring the row vector of shadow prices, the matrix of coefficient rows, and the objective coefficients. The first two components of the dual constraint, $wx^1 \geq py^1$ and $wx^2 \geq py^2$, state that the prices render the two activities unprofitable. Rewriting, $w/p \geq y^1/x^1$ and $w/p \geq y^2/x^2$. By assumption that the first firm is more productive and because w is minimized, the first dual constraint is binding, $w/p = y^1/x^1$. In other words, *the real wage rate equals the highest productivity* and, therefore, this activity would break even.

There is an interesting connection between shadow prices and competitive markets. Without loss of generality, program (1) has been set up (by inclusion of coefficient y^1 in the objective function) such that the third component of the dual constraint, $py^1 = y^1$, normalizes the price system $(w\ p)$ such that $p = 1$. Hence $w = y^1/x^1$. The second, less productive, activity would be unprofitable under these prices. In other words, if shadow prices prevail and entrepreneurs are profit maximizers, they would select the optimal activity to produce output. The solutions to the profit maximization problems are not unique, but there exists a combination of solutions which is consistent with equilibrium; see ten Raa and Mohnen (2002).

In the primal program (1), the less productive activity is suppressed by setting $\theta_2 = 0$ and, therefore, $\theta_1 \leq 1$ and the maximum value is $c = 1$. Firm 1 cannot expand its output. Next consider firm 2. In program (1), in the right sides of the constraints, superscripts 1 are replaced by 2. The first two dual constraints, $wx^1 \geq py^1$ and $wx^2 \geq py^2$, remain the same, as does the conclusion that the first activity would be adopted to produce output. The maximum expansion factor equals the ratio of the highest productivity to the actual productivity, $c = (y^1/x^1)/(y^2/x^2)$. For example, if this number is 1.25,

potential output of firm 2 exceeds actual output by 25%. Conversely, actual output is only 80% of potential output. Firm 1, however, produces 100% of its potential output. The efficiency of firm 1 is 100% and the efficiency of firm 2 is 80%. Here efficiency is defined as the inverse expansion factor, $(y^2/x^2)/(y^1/x^1)$ for firm 2 and $(y^1/x^1)/(y^1/x^1)=1$ for firm 1. Efficiency is the performance measure. *Efficiency is equal to the ratio of actual to optimal productivity.* In the single input-single output case with constant returns to scale, introduced in this section, efficiency must be technical efficiency. However, in more general settings, the inverse expansion factor of efficiency will also encompass allocative efficiency, as we will see in Sects. 2 and 4.

2 Multiple Input-Single Output Production

In the bulk of the economic literature, including macro-economics, there are multiple inputs, such as labor and capital, but a single output. The inputs, x^1 for firm 1 and x^2 for firm 2, turn vectors and the input price will be represented by row vector w. The previous set-up is maintained and the extension to more than two firms is straightforward. However, because of the multiplicity of inputs, several activities may now be activated when a firm maximizes output given its input vector, x. The potential output given an input vector is a scalar, the value of a function, $y=F(x)$. This is the reduced form of program (1) with y equal to scalar y^1c and x equal to vector x^1. Mind that potential output y is the product of actual output and the expansion factor. F is called the *production function*. To define productivity as an output/input ratio, we must aggregate the input components, if only because division by a vector is impossible. The way to do this is intuitive, making use of a well-known property of Lagrange multipliers, namely that they measure the gain in output per unit of input. The rate of potential output with respect to input k is given by w_k, the shadow price of the k^{th} component of the constraint $\sum_i x^i\theta_i \leq x$. The productivity of input k is shadow price w_k. This is output per unit of input. Now the problem is that a unit of input is arbitrary. For example, sugar can be measured in kilograms or in metric pounds. Using the latter, a unit has half the size of the former, the number of units is doubled, and the shadow price is halved. We must aggregate across inputs in a way that is not sensitive with respect to the units of measurement. The way to do this is to first aggregate productivity over the units of the same input, k. The contribution to output of input k is w_kx_k and in this product, the two effects of taking metric pounds instead of kilograms cancel. Summing over inputs k, the value of the dual program is obtained. However, by the main theorem

of linear programming, the value of the dual program is equal to the value of the primal program, potential output y. The aggregate output/input ratio, $y/\sum_i wx^i$, is thus unity. The reason for this peculiar limitation is that output and input are different commodities; there is no common denominator. This is the economic problem of value and the classical solution is to express output in terms of resource contents, like labor values. Then, indeed, the output/input ratio is bound to be one.

Yet this framework is useful, because productivity levels are determined relative to a base economy, a base year. We do observe changes in the output/input ratio over time. For example, if the productive firm in Sect. 1, firm 1, increases output in the next period, then $w=y^1/x^1$ remains valid, hence productivity w increases. This argument is extendable to the multi-input case. Dropping firm indices, productivity growth of input k is $\dot{w}_k$, where the dot stands for the derivative with respect to time. This, again, is sensitive with respect to the unit of measurement. However, aggregating across inputs, weighing by the units of inputs, $\sum_k \dot{w}_k x_k$, the sensitivity gets lost, because $\dot{w}_k x_k = \frac{\dot{w}_k}{w_k} w_k x_k$, in which the ratio is a growth rate while the subsequent product was already seen to be insensitive with respect to the unit of measurement. It is also customary to express the change in the output/input ratio as a growth rate, by dividing by the level, $\sum_k w_k x_k$. In short, the output/input ratio grows at the rate

$$TFP = \sum\nolimits_k \dot{w}_k x_k / \sum\nolimits_k w_k x_k. \tag{2}$$

Expression (2) is called *total factor productivity growth*. *TFP* is the most prominent performance measure. The expression can be rewritten as a weighted average of the factor productivity growth rates, $\dot{w}_k/w_k$, with weights $w_k x_k / \sum w_k x_k$. These weights sum to one.

This direct approach from Lagrange multiplier-based input productivities to total factor productivity growth can be related to the Solow residual approach. Recall that the values of the primal and dual programs match, $py = wx$, where the right-hand side is the product of row vector w and column vector x, and that we normalized $p=1$. Differentiating totally, $\dot{w}\,x = \dot{y} - w\,\dot{x}$ and, therefore, expression (2) equals

$$TFP = (p\,\dot{y} - w\,\dot{x})/py. \tag{3}$$

Expression (3) is called the Solow residual; see ten Raa (2008), Sect. 7, and the references given there. Solow (1957) modeled technical change by letting the production function depend on time,

$$y = F(x,t). \tag{4}$$

Differentiating production function (4) with respect to time, indicating partial derivatives by subscripts, $\dot{y} = F'_x \dot{x} + F'_t$ or

$$(\dot{y} - F'_x \dot{x})/y = F'_t/F, \tag{5}$$

Now, if inputs are rewarded according to their marginal products, $w = pF'_x$, then the left-hand sides of Eqs. (3) and (5) match, and, therefore, the Solow residual (3) reduces to F'_t/F, i.e., *technical change*. This condition is fulfilled if the input prices are the shadow prices of the program that maximizes output subject to technical feasibility. The production possibility set, $\{(x, y): y \leq F(x, t)\}$, is the set which is either spanned by the observed input-output pairs or postulated by some functional form of function F. This distinction corresponds with nonparametric and parametric performance measurement. The first underpinning, by observed input-output pairs, is more fundamental, as the second underpinning, by a production function, can be shown to be generated by a distribution of input-output pairs, where the distribution represents the capacities of the activities. Houthakker (1955) demonstrated this for the Cobb–Douglas function, $Y = AK^{\alpha}L^{\beta}$, where K and L are inputs, Y is output, and A, α, and β are parameters with $\alpha + \beta < 1$, meaning there are decreasing returns to scale. The returns to scale decrease because of constraining third input, as will be explained next. Output notation Y is customary in the Cobb–Douglas literature. Moreover, we may now reserve y for full capacity output.

An *activity* is a pair of proportionate inputs and an output. The assumption of input proportionality facilitates normalization of the activity by the output to $(k, l; 1)$, with $k = K/Y$ and $l = L/Y$ fulfilling $Ak^{\alpha}l^{\beta} = 1$. The activities can be parameterized by one input, e.g., k. Then $l = (Ak^{\alpha})^{-1/\beta}$ and, therefore, the technology set of activities is $\{(k, (Ak^{\alpha})^{-1/\beta}; 1): k > 0\}$. Each activity can be run with intensity s_k. Total output will be $\int s_k \mathrm{d}k$, where the integral is taken over the positive numbers. The constraints are $\int s_k k \mathrm{d}k \leq K$ and $\int s_k l \mathrm{d}l \leq L$, where K and L are the factor endowments. However, Houthakker (1955) assumes there is a capacity constraint for each activity. A fixed input causes the capacity constraint. The fixed input is different than the variable inputs, capital, and labor. Houthakker (1955) suggests entrepreneurial resources. The distribution of entrepreneurial resources (i.e., of the capacity constraint) across activities $(k, l; 1)$ is considered to be given and denoted by $y(k, l)$. This distribution need not be concentrated on a frontier-like $\{(k, l): Ak^{\alpha}l^{\beta} = 1\}$. Some activities may dominate others, with

both components of (k, l) smaller. Yet a dominated activity may be run, because the superior activity, like all activities, has a capacity constraint. Activities can be run with intensities $0 \leq s(k, l) \leq y(k, l)$. Subject to the factor constraints $\iint s(k,l)k\mathrm{d}k\mathrm{d}l \leq K$ and $\iint s(k,l)l\mathrm{d}k\mathrm{d}l \leq L$, we maximize output $\iint s(k,l)\mathrm{d}k\mathrm{d}l$. This is an infinite-dimensional linear program, with a continuum of variables $s(k, l)$. Denote the shadow prices of the two factor constraints by r and w, respectively. By the phenomenon of complementary slackness, unprofitable activities, with unit cost $rk + wl > 1$, are not run, $s(k, l) = 0$. By the same argument, profitable activities, with unit cost $rk + wl < 1$, are run at full capacity, $s(k, l) \leq = y(k, l)$. Activities which break even, $rk + wl = 1$, have activity $0 \leq s(k, l) \leq y(k, l)$, but since the set of such activities has measure zero, we may set $s(k, l) = y(k, l)$. It follows that inputs and output are $K = \iint\limits_{rk+wl\leq 1} y(k,l)k\mathrm{d}k\mathrm{d}l$, $L = \iint\limits_{rk+wl\leq 1} y(k,l)l\mathrm{d}k\mathrm{d}l$, and $Y = \iint\limits_{rk+wl\leq 1} y(k,l)k\mathrm{d}k\mathrm{d}l$, respectively. The implicit assumption is that all factor input can be fully employed. There must be activities with factor intensity k/l below endowment ratio K/L and activities with factor intensity above the endowment ratio.

The three expressions, for inputs K and L and output Y, are interrelated by the two shadow prices r and w. The idea of Houthakker (1955) is to use the first two expressions to solve for r and w in terms of K and L. Substitution of the results in the third expression yields output as function of the inputs. Houthakker (1955) carries out this calculation for the capacity distribution with Pareto density function, $y(k, l) = \mu k^{\kappa-1} l^{\lambda-1}$, where μ, κ, and λ are positive constants. The result is $Y = AK^{\alpha}L^{\beta}$ with $\alpha = \kappa(\kappa + \lambda + 1)$, $\beta = \lambda(\kappa + \lambda + 1)$ and A a positive constant depending on μ, κ, and λ. In other words, a Pareto capacity distribution yields a Cobb–Douglas production function. This is Houthakker's Theorem. At the micro-level, activities have fixed input-output ratios—it takes given amounts of labor to operate given machinery and equipment—but a change in resources, such as the inclusion of the East German labor force in the year 1989, is accommodated by the activation of new activities and the deactivation of some incumbent activities. Reallocations of resources across activities manifest as substitutions.

The capacity distribution is not concentrated on a single isoquant in input space. Both k and l can be bigger, less efficient. In solving the output maximization, smaller input combinations are activated, but only to full capacity. Residual inputs are employed by more input intensive activities. The capacity constraints thus yield decreasing returns to scale. Indeed, the Cobb–Douglas function has exponents summing to a number less than one.

Houthakker's activity foundation of neoclassical production functions works only if returns to scale are decreasing.

Clearly, different capacity distributions for the activity levels will generate different production functions. Houthakker (1955) has generated a stream of theoretical and applied research. The bulk of this literature features a lower dimension, with only one variable input, namely labor, and again one fixed output, which is now capital. In this one fixed-one variable input framework, Levhari (1968) found the capital distribution for which total output is a CES function of the total fixed input (capital) and the total variable input (labor) and showed it encompasses the Cobb–Douglas function. Muysken (1983) has consolidated the Cobb–Douglas, CES, and VES functions by showing they are all generated by beta distributions, with alternative parametrizations. Two books on the distribution approach to production are Johansen (1972) and Sato (1975). In this literature, activities have fixed input-output proportions and capacity constraints explain the existence of inefficient activities. Increases in levels of inputs prompt the activation of less efficient activities, in Ricardian style. The law of one price yields rents to the more efficient activities. The activation of different activities prompts different proportions between the input totals and the output. Substitution is considered a symptom of the change in the range of active activities (run with positive intensity).

3 Single Input-Multiple Output Production

In classical economics, labor is the only factor input. All other inputs are produced commodities, also called intermediate inputs. Production output is used to fulfill intermediate demand and final demand, where the latter is defined residually, as the difference between output and intermediate input. Production output is also called gross output; similarly, final demand is also called net output. In standard input-output analysis, each output has a single technique to produce it. Assuming constant returns to scale, the input of commodity j, $j = 1, \ldots, n$, per unit of output, is denoted by the input vector $(a_{1j}, \ldots, a_{nj}, l_j)$, of which the components represent the n intermediate inputs and the factor input (labor), respectively. If these unit input requirements are constant and fixed, they cannot be reduced and, therefore, are necessarily efficient (actual and optimal production coincide). If, however, there is a set of input vectors for each product j, there is room to reallocate labor between alternative techniques, which may save labor or, alternatively, increase output. This would increase the output/input ratio from

actual productivity to optimal productivity. The ratio of the two would be efficiency. A deep result states that the optimal input vectors, one for each product, are independent of the composition of final demand. This is the substitution theorem, but for an obvious reason also called non-substitution theorem, which goes back to Samuelson (1951). The proof of the theorem has a long history, in which details have been worked out and minor flaws eliminated. This culminated in a proof based on the efficiency program of maximizing the expansion factor of a some net output vector, determining the optimal input vectors, one for each product, and showing that this combination of input vectors remains optimal when the net output vector is replaced by another one (ten Raa 1995).

The substitution theorem yields an all-purpose optimal technology, featuring one technique for each product. Given any net output vector, one can compare the optimal labor input to the actual labor input. The ratio is the efficiency of the economy.

4 Multiple Input-Multiple Output Production

The determination of efficiency is simple in the single output case: One maximizes output given the inputs and in the single input case, one can minimize the input given the output. A mechanical extension to the multiple input-multiple output world would be to expand the output vector while preserving its component proportions. This procedure, however, presumes that the mix of outputs should not be changed and is optimal. Yet it is a useful procedure and I will detail it and discuss its merits. The fundamental paper of this approach is Debreu's (1951) now classic "The Coefficient of Resource Utilization," which will be discussed first.

The economy comprises m consumers with preference relationships $\succsim_i$ and observed l-dimensional consumption vectors $y^i (i = 1, \ldots, m)$, where l is the number of commodities.[1] Z is the set of possible l-dimensional input vectors (*net* quantities of commodities consumed by the whole production sector during the period considered), including the observed one, z. A combination of consumption vectors and an input vector is *feasible* if the total sum—the economy-wide *net* consumption—does not exceed the vector of

[1] I stick to the performance literature notation of (factor) inputs x, (consumed) outputs y, and intermediate inputs z. In the general equilibrium literature, including Debreu (1951), the notation is (factor) inputs z, (consumed) outputs x, and intermediate inputs y.

utilizable physical resources, l-dimensional vector x.[2] Vector x is assumed to be at least equal to the sum of the observed consumption and input vectors, ensuring the feasibility of the latter.

The set of net consumption vectors that are at least as good as the observed ones is

$$B = \left\{ \sum y^{i'} : y^{i'} \succsim_i y^i, \quad i = 1, \ldots, m \right\} + Z. \tag{6}$$

The symbol B stands for "better" set. The minimal resources required to attain the same levels of satisfaction that come with x^i belong to $B^{\min}$, the south-western edge or subset of elements z' that are minimal with respect to $\geq$.[3] Assume that preferences $\succsim_i$ are convex and continuous, and that production possibilities form a convex and closed set, then the separating hyperplane theorem yields a supporting price row vector $p(x') > 0$ (all components positive) such that $x'' \in B$ implies $p(x')x'' \geq p(x')x'$. The *Debreu coefficient of resource utilization* is defined by

$$\rho = \max_{x'} \{ p(x')x'/p(x')x : x' \in B^{\min} \}. \tag{7}$$

Coefficient ρ measures the distance from the set of minimally required physical resources, $x' \in B^{\min}$, to the utilizable physical resources, x, in the metric of the supporting prices (which indicate welfare indeed). Debreu (1951, p. 284) shows that the distance or the max in (7) is attained by

$$x' = \rho x \in B^{\min}. \tag{8}$$

In other words, the Debreu coefficient of resource utilization is the smallest fraction of the actually available resources that would permit the achievement of the levels of satisfaction that come with x^i. Coefficient ρ is a number between zero and one, the latter indicating full efficiency. In modern terminology, this result means that ρ is the *input-distance function*, determined by the program

$$\rho = \min_r \left\{ r : \sum y^{i\prime} + z' \leq rx, y^{i\prime} \succsim_i y^i, z' \in Z \right\}. \tag{9}$$

[2]For example, if the last commodity, l, represents labor, and this is the only nonproduced commodity, then $x = Ne_l$, where N is the labor force and e_l the l-th unit vector.

[3]By convention, this vector inequality holds if it holds for all components.

5 The Farrell Efficiency Measure

Another classic paper is Farrell (1957), which decomposes efficiency in technical efficiency and price efficiency. Here, technical efficiency is measured by the reduced level of proportionate inputs (as a percentage between 0 and 100) such that output is still producible. Price efficiency is the fraction of the value of an input vector with possibly different proportions (but the same output) to the value of the efficient input vector with the given proportions. Farrell (1957) notes the similarity between his technical efficiency and the Debreu coefficient of resource utilization: Both the Farrell technical efficiency measure and the Debreu coefficient of resource utilization are defined through proportionate input contractions, but the analogy is sheer formality and even confusing at a conceptual level. The analogy suggests that Farrell takes the Debreu coefficient to measure technical efficiency and augments it with a reallocative efficiency term, thus constructing a more encompassing overall measure. However, it is the other way round; the sway of the Debreu coefficient is far greater than that of Farrell's efficiency measure. Farrell's price efficiency measure is a partial (dis)equilibrium concept, conditioned on prices. It takes into account the cost reduction attainable by changing the mix of the inputs, given the prices of the latter.

The Debreu coefficient, however, is a general (dis)equilibrium concept. It measures the technical and allocative inefficiency in the economy given only its fundamentals: resources, technology, and preferences. Prices are derived and enter the definition of the Debreu coefficient, see (2). Debreu (1951) then *proves* that the coefficient can be freed from these prices, by Eq. (8) or non-linear program (9). The prices are implicit, supporting the better set in the point of minimally required physical resources. The Debreu coefficient measures technical *and* allocative inefficiency, both in production and in consumption, solving the formidable difficulty involved in assessing prices, referred to by Charnes et al. (1978, p. 438). Farrell refrains from this, restricting himself to technical efficiency and price-conditioned allocative efficiency, which he calls price efficiency.

The formal analogy between the Debreu coefficient and the Farrell measure of technical efficiency prompted Zieschang (1984) to coin the phrase "Debreu-Farrell measure of efficiency," a term picked up by Chakravarty (1992) and Grifell-Tatjé et al. (1998), but this practice is confusing. Debreu's coefficient of resource allocation encompasses both Farrell's technical efficiency and his price efficiency measures and frees the latter from prices. On top of this, Debreu's coefficient captures consumers' inefficiencies. The confusion persists. Färe et al. (2002) speak of the "Debreu-Farrell

measure of technical efficiency." A recent review of Farrell's contribution states

> (Debreu) worked only from the resource cost side, defining his coefficient as the ratio between minimised resource costs of obtaining a given consumption bundle and actual costs, for given prices and a proportional contraction of resources. Førsund and Sarafoglou (2002, footnote 4)

However, Debreu (1951) calculates the resource costs *not* of a given consumption bundle, but of an (intelligently chosen) Pareto equivalent allocation. (And the prices are not given, but support the allocation.) It is true, however, that the Debreu measure would become applicable if the aggregated consumption bundle can be considered given. Ten Raa (2008) demonstrates that this approach is doable and that it is exact if the preferences are Leontief.

6 The Debreu–Diewert Coefficient of Resource Utilization

Diewert (1983) had the idea that Leontief preferences remove misallocations between consumers as a source of inefficiency. The consequent coefficient of resource utilization yields a more conservative estimate of inefficiency than Debreu's coefficient resource of utilization. Ten Raa (2008) shows that Leontief preferences not only separate production efficiency from consumption efficiency, but also solve an aggregation problem: The Leontief preferences may vary between consumers, with different preferred consumption bundle proportions, but information of this preference variation need not be given. This useful fact is explained now.

Leontief preferences $\succsim_i$ with nonnegative bliss point y^i are defined for nonnegative consumption vectors by $y'' \succsim_i y'$ if $\min y''_k/y_k \geq \min y'_k/y_k$ where the minimum is taken over commodities $k = 1, \ldots, l$. If so, the consumption term in better set (6) fulfills (ten Raa, 2008)

$$\left\{\sum y^{i\prime} : y^{i\prime} \succsim_i y^i,\ i = 1, \ldots, m\right\} = \left\{y' : y' \geq \sum y^i\right\}. \tag{10}$$

Equation (10) shows that "more is better" at the micro-level if and only if "more is better" at the macro-level. Equation (10) is a perfect aggregation result. One might say that if preferences are Leontief with varying bliss points (according to the observed consumption baskets), there is a social

welfare function. The better set is freed from not only preferences, $\succeq_i$, but also individual consumption baskets, y^i. Only *aggregate* consumption is required information.

This result creates the option to determine the degree of efficiency in terms of output. If the production set X features the impossibility to produce something from nothing and constant returns to scale, then $\gamma = 1/\rho$ transforms the input-distance function program (9) into the *output-distance function* program

$$1/\rho = \max\{c : c \sum y^i + z' \leq x, z' \in Z\}. \tag{11}$$

Output-distance program (11) determines the expansion factor and potential consumption, i.e., net output. The ratio of actual output to potential output is equal to efficiency, the Debreu–Diewert coefficient of resource utilization, ρ. This has been applied and analyzed, including decompositions in different inefficiency components, for various economies.

Ten Raa and Mohnen (2001) evaluate the gains from free trade between the European and Canadian economies. The results show that bilateral trade liberalization would multiply the trade volume and let Canada, which is a small economy, to specialize in a few sectors. Perfect competition and free trade together will result in the expansion factors of 1.075 for Europe and 1.4 for Canada, while without free trade the economies expand to 1.073 and 1.18, respectively. The gains of free trade are evaluated at 0.2% for Europe and 22% for Canada. Sikdar et al. (2005) apply a similar model for measuring the effects of freeing bilateral trade between India and Bangladesh. The study was conducted against the background that Bangladesh was about to join the South Asian Association for Regional Cooperation (SAARC, established in 1985), in which India participated from the very beginning. Using the linear program version of the model, the authors locate comparative advantages in both economies and contrast them with the observed trade pattern. While the patterns are generally comparable, there are notable differences for some products. For example, it turns out that although India is an exporter of "Livestock, fishing, forestry" and "Other food products," the free trade model suggests that these should be import products for India. While on its own, each economy's expansion factor equals 1.37, the introduction of free trade would increase it to 1.43 for India and 1.97 for Bangladesh. This means that the potential gains of free trade for these two countries are 6% and 60%. Similarly to the previous paper, a small economy—Bangladesh—has much more to gain by joining the free trade agreement with a large economy. Ten Raa (2005) evaluates the contribution of

international trade, disentangling trade efficiency from domestic efficiency and splits the domestic efficiency of the economy into X-efficiency and allocative efficiency.

Another interesting decomposition of efficiency is provided by Cella and Pica (2001), who use a convex piecewise linear envelopment of the observed data (DEA) to disentangle sectoral inefficiencies in five OECD countries, Canada, France, Denmark, Germany, and the UK, into internal sectoral inefficiencies and inefficiencies imported from other sectors through the price distortion of intermediate product prices. These imported inefficiencies are also called "spillovers" from other sectors. The study shows that inefficiency spillovers are empirically relevant in all sectors of the five considered countries.

Amores and ten Raa (2014) distinguish three levels of production efficiency of the Andalusian economy: a firm level, an industry level, and the economy level. *Firm level* efficiency measures the potential productivity gains (i.e., output/input ratios) that arise if the firm could choose to use production techniques of other firms from the same industry. (However, intellectual property rights may prevent this.) *Industry efficiency* measures the gains that can be achieved by pooling all the vectors of inputs and outputs of the firms that belong to this industry and reallocating production within the industry to maximize the total output value of the industry. Finally, the total *efficiency of the economy* measures the gains that can be achieved by the economy if there were no barriers to reallocation of inputs and outputs across firms and industries. Based on the results from these three problems, one can distinguish *industrial organization efficiency* and *industrial specialization efficiency*. The former captures the efficiency gains achieved by reorganization within industries, if each industry starts to produce a more valuable (i.e., efficient) output mix. The latter captures the additional efficiency that can be achieved by re-specialization of the output mix of the economy.

7 Interrelation Between the Productivity and Efficiency Measures

Productivity growth, measured by the Solow residual (3), and efficiency, measured by the Debreu–Diewert coefficient of resource utilization (11), can be interrelated.

Productivity is output per input. For an economy, input are the resources and output is the final consumption. Input x and output y are multi-dimensional. Denote the production possibility set at time t, the set of all pairs (x, y)

such that x can produce y at time t by P^t, the so-called *production possibility set*. Following Eq. (9) the input-distance function is

$$D(x, y, t) = \min\{r : (rx, y) \in P^t\}. \tag{12}$$

Input distance r is a number between zero and one. If $r = 1$, input cannot be contracted, is on the frontier of the production possibility set, and is efficient. If $r < 1$, input can be contracted, is not on the frontier, and is inefficient. An increase in the input distance signals an increase in efficiency. *Efficiency change* is the relative change in input-distance function (12) with a dot representing time derivative:

$$EC = \dot{D}/D. \tag{13}$$

The distance to the frontier may grow without any change in input x or output y, simply because the frontier shifts out. This shows a decrease in the input distance. *Technical change* is minus the relative partial derivative of the input-distance function with respect to time, i.e., keeping input x and output y fixed:

$$TC = -D'_t/D. \tag{14}$$

To relate these efficiency change and technical change to the single output Solow residual analysis, we must replace Solow's implicit assumption that output is related to input by the production function, (4), by the more general relationship

$$y = D(x, y, t)F(x, t), \tag{15}$$

where potential output is reduced to actual output. Differentiating Eq. (15) with respect to time, $\dot{y} = \dot{D}F + D(F'_x \dot{x} + F'_t)$ or, dividing by expression (15),

$$(\dot{y}/y - F'_x \dot{x}/F) = \dot{D}/D + F'_t/F, \tag{16}$$

The left-hand side of formula (16) features total factor productivity, see Equation with y 1-dimensional and p canceling out, and the right-hand side features efficiency change (13) plus technical change. The last term is indeed consistent with Eq. (14), as output $y = D(x, y, t)F(x, t)$ and partial differentiation with respect to time yield $D'_t F + DF'_t = 0$. Summarizing,

$$TFP = EC + TC = \dot{D}/D - D'_t/D, \tag{17}$$

where the second equality holds term by term. Expression (17) holds for multi-output production with, however, constant returns to scale. Ten Raa (2008) proves that the efficiency change term is measured by the growth rate of the Debreu–Diewert coefficient of resource utilization and the technical change term by a generalized Solow residual of net frontier output growth evaluated at the supporting price vector.

In applied work, time is in discrete periods and the main performance measure that accommodates this is the Malmquist productivity index (Caves et al. 1982). Its derivation is as follows. The first term on the right-hand side of Eq. (17) is the total derivative of input distance $D(x, y, t)$ and the last term subtracts the third partial derivative. What remains are the first two partial derivatives,

$$TFP = \frac{\partial \ln D(x,y,t)}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial \ln D(x,y,t)}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}. \tag{18}$$

In discrete time expression (18) is a local approximation to

$$\ln D(x^{t+1}, y^{t+1}, \bullet) - \ln D(x^t, y^t, \bullet) = \ln \frac{D(x^{t+1}, y^{t+1}, \bullet)}{D(x^t, y^t, \bullet)}. \tag{19}$$

Evaluating this expression at t and $t+1$, taking the average of the two logarithms and exponentiating, one obtains the standard expression of the Malmquist productivity index:

$$TFP = \left[\frac{D(x^{t+1}, y^{t+1}, t)}{D(x^t, y^t, t)}\frac{D(x^{t+1}, y^{t+1}, t+1)}{D(x^t, y^t, t+1)}\right]^{1/2}. \tag{20}$$

The explicit price information in the Solow residual (3) has been replaced by implicit shadow price information, derived from the shape of the frontier; see Coelli and Rao (2001). The Malmquist productivity index assumes constant returns to scale. The decomposition of the Malmquist index into technical change and efficiency change, see Eq. (17), is straightforward; see Färe et al. (1994).

The Malmquist productivity index is popular because of its simplicity. Moreover, it can be bridged with other important TFP growth indices. The Törnqvist productivity index is defined by the discrete-time approximation of (3) with value weights $w_k x_k / wx$ and $p_k y_k / py$ approximated by their arithmetic averages between periods t and $t+1$ and growth rates $\dot{x}_k / x_k$ and $\dot{y}_k / y_k$ approximated by the changes in the logs of x and y between periods t and

$t+1$. Caves et al. (1982) have shown that the Malmquist productivity index becomes a Törnqvist productivity index provided that the distance functions are of translog form with identical second-order coefficients and that the prices support cost minimization and profit maximization. The Fisher productivity index is also defined by a discrete-time approximation of (3), with the changes in the logs of x and y now evaluated at the prices in periods t and $t+1$ separately and then averaged arithmetically. Färe and Grosskopf (1996) have proved that the Malmquist productivity index approximates the Fisher productivity index under the assumption of profit maximizing behavior. Balk (2008) reviews comprehensively, including non-constant returns to scale.

A defect of the Malmquist, Törnqvist, and Fisher indices is that they are not transitive. The changes from periods t to $t+1$ and from periods $t+1$ to $t+2$ do not add to the change from periods t to $t+2$. A necessary and sufficient condition for transitivity is that the index between periods can be written as a ratio of values of a function evaluated in the two periods. This property is fulfilled for the efficiency change component of productivity growth, but not for the technical change component, unless technical change is Hicks neutral. However, Balk and Althin (1996) shows that a modification of the Malmquist index, averaging out between firm observations, is transitive.

8 Conclusion

The key concept in performance analysis is productivity, which is the output/input ratio. Both output and input are aggregates. The appropriate weights are shadow prices of the program that determines potential output. The latter is based on observed input-output pairs or a production function, corresponding with nonparametric and parametric performance analysis, respectively. Parametric performance analysis can be conceived as nonparametric performance analysis with an appropriate distribution of observations. Hence nonparametric analysis is more fundamental. Replacing output by potential output, productivity becomes optimal productivity. The ratio of actual productivity to optimal productivity is equal to efficiency. Performance may increase because of efficiency change, technical change, scale economies, or changes in the production environment. Technical change is a change in optimal productivity. All this can be grounded in economic theory, where optimality is defined in terms of consumer preferences. If consumers have Leontief preferences, with consumptions bundles

preferred to be in fixed proportions, which may vary between consumers, then performance analysis is freed from micro-consumer data requirements and shadow prices can be determined on the basis of production data and the proportions of final demand. Moreover, then the efficiency is measured by Debreu's coefficient of resource utilization and technical change by the Solow residual of net frontier output growth.

Acknowledgements I am grateful to a referee for detailed criticism that prompted numerous improvements.

References

Amores, A.F., and T. ten Raa. 2014. Firm efficiency, industry performance and the economy: Three-way decomposition with an application to Andalusia. *Journal of Productivity Analysis* 42 (1): 25–34.

Balk, B. 2008. *Price and quantity index numbers*. Cambridge: Cambridge University Press.

Balk, B.M., and R. Althin. 1996. A new, transitive productivity index. *Journal of Productivity Analysis* 7 (1): 19–27.

Blackorby, C., and R.R. Russell. 1999. Aggregation of efficiency indices. *Journal of Productivity Analysis* 12 (1): 5–20.

Caves, D.W., L.R. Christensen, and W.E. Diewert. 1982. The economic theory of index numbers and the measurement of input, output and productivity. *Econometrica* 50 (6): 1393–1414.

Cella, G., and G. Pica. 2001. Inefficiency spillovers in five OECD countries: An interindustry analysis. *Economic Systems Research* 13 (4): 405–416.

Chakravarty, Satya R. 1992. Efficiency and concentration. *Journal of Productivity Analysis* 3: 249–255.

Charnes, A., W.W. Cooper, and E. Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2: 419–444.

Coelli, T., and D.S.P. Rao. 2001. Implicit value shares in Malmquist TFP index numbers. CEPA Working Paper No. 4/2001, University of New England, Armidale, Australia.

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

Diewert, W. Erwin. 1983. The measurement of waste within the production sector of an open economy. *Scandinavian Journal of Economics* 85 (2): 159–179.

Färe, R., and S. Grosskopf. 1996. *Intertemporal production frontiers: With dynamic DEA*. Boston: Kluwer Academic Publishers.

Färe, R., S. Grosskopf, B. Lindgren, and P. Roos. 1994. Productivity developments in Swedish hospitals: A Malmquist output index approach. In *Data Envelopment Analysis: Theory, Methodology and Applications*, ed. A. Chames, W.W. Cooper, A. Lewin, and L. Seiford. Boston: Kluwer Academic Publishers.

Färe, R., S. Grosskopf, and V. Zelenyuk. 2002. Finding common ground: Efficiency indices. UPEG Working Paper 0305, Presented at the North American Productivity Workshop at Union College, Schenectady, NY.

Farrell, M.J. 1957. The measurement of productive efficiency. *Journal of Royal Statistical Society* 120 (3): 253–290.

Førsund, F.R., and N. Sarafoglou. 2002. On the origins of data envelopment analysis. *Journal of Productivity Analysis* 17: 23–40.

Grifell-Tatjé, E., C.A.K. Lovell, and J.T. Pastor. 1998. A quasi-Malmquist productivity index. *Journal of Productivity Analysis* 10 (1): 7–20.

Houthakker, H.S. 1955. The Pareto distribution and the Cobb-Douglas production function in activity analysis. *Review of Economic Studies* 23 (1): 27–31.

Johansen, L. 1972. *Production functions: An integration of micro and macro, short run and long run aspects*. Amsterdam, The Netherlands: North-Holland.

Levhari, D. 1968. A note on Houthakker's aggregate production function in a multifirm industry. *Econometrica* 36 (1): 151–154.

Milgrom, P.R. 1981. Rational expectations, information acquisition, and competitive bidding. *Econometrica* 49 (4): 921–943.

Muysken, J. 1983. Transformed beta-capacity distributions of production units. *Economics Letters* 11 (3): 217–221.

Samuelson, P.A. 1951. Abstract of a theorem concerning substitutability in open Leontief models. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans, 142–146. New York: Wiley.

Sato, K. 1975. *Production functions and aggregation*. Amsterdam, The Netherlands: North-Holland.

Sikdar, C., D. Chakraborty, and T. ten Raa. 2005. A new way to locate comparative advantages of India and Bangladesh on the basis of fundamentals only. In *Essays on international trade, theory and policy for the developing countries*, ed. R. Acharyya, 169–197. Kolkata: Allied Publishers.

Solow, R.M. 1957. Technical change and the aggregate production function. *The Review of Economics and Statistics* 39 (3): 312–320.

ten Raa, T. 1995. The substitution theorem. *Journal of Economic Theory* 66 (2): 632–636.

ten Raa, T. 2005. *The economics of input-output analysis*. Cambridge: Cambridge University Press.

ten Raa, T. 2008. Debreu's coefficient of resource utilization, the Solow residual, and TFP: The connection by Leontief preferences. *Journal of Productivity Analysis* 30 (3): 191–199.

ten Raa, T., and P. Mohnen. 2001. The location of comparative advantages on the basis of fundamentals only. *Economic Systems Research* 13 (1): 93–108.

ten Raa, T., and P. Mohnen. 2002. Neoclassical growth accounting and frontier analysis: A synthesis. *Journal of Productivity Analysis* 18 (2): 111–128.

Zieschang, K.D. 1984. An extended Farrell technical efficiency measure. *Journal of Economic Theory* 33 (2): 387–396.

R&D, Innovation and Productivity

Pierre Mohnen

1 Introduction

In the short run, labor productivity (output per hour worked), capital productivity (output per unit of capital stock) or total factor productivity (TFP) (a weighted sum of outputs divided by a weighted sum of inputs) varies over the business cycle because of inflexibilities of various sorts: hiring and firing costs, labor regulations, time to build or adjustment costs leading to variations in capacity utilization. In the long run, however, changes in technology alter technical coefficients—the amount of a certain input needed per unit of output—augmenting the marginal productivity of certain factors of production or saving on some of them and thereby affect TFP.

For a long time, technological change was considered as exogenous or simply measured by a time trend. In the last 50 years, various theories have been developed to try and explain the phenomenon of technological change and its impact on economic growth. Various indicators have been collected in order to better understand how it occurs and what effect it has on the level and the growth rate of TFP.

This chapter goes over various technological indicators—R&D expenditure, patents, patent citations, innovation expenditure, the share of innovative sales, count data of innovations and various measures of purchased technologies—pointing out their strengths and weaknesses and the

P. Mohnen (✉)
Maastricht University and UNU-MERIT, Maastricht, The Netherlands
e-mail: mohnen@merit.unu.edu

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_4

consequent measures of caution to be taken when using these data for economic analysis. It briefly explains the theoretical link between innovation and productivity growth and then compares the estimated magnitudes of that relationship using the different innovation indicators.

The rest of the paper is organized as follows. First, it reviews the most frequently used indicators of technology and discusses their pros and cons. It then examines how they have been used to explain changes in productivity, what econometric challenges are posed by each indicator and what have been the major results obtained. It concludes with some reflections on the merits of indicators and on the state of knowledge regarding the link between innovation and TFP.

2 Technological Indicators

It is useful to start with a description of the data sources available to study the link between innovation and productivity. I shall cover in detail three types of data, which are available in most countries: R&D surveys, patent statistics and innovation surveys. I shall say a few words about other data sources, less frequently used or only available sporadically in a limited number of countries.[1]

2.1 Research and Development Surveys

According to the Frascati Manual (OECD 2015), "Research and experimental development (R&D) comprises creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications." It excludes things like routine testing, the analysis of materials, feasibility studies, routine software development and general purpose data collection. R&D can be decomposed into basic research, applied research and experimental development. It can be performed and/or funded by the business enterprise sector, government, higher education and private non-profit organizations.

Starting with the pioneering work of Griliches and Mansfield in the late 1950s and early 1960s (Griliches 1964; Mansfield 1965), a large literature has developed in which R&D expenditures are considered as investments in a

[1]For a more extended discussion on innovation indicators, see Kleinknecht (2002), Hagedoorn and Cloodt (2003), Gault (2010, 2013), and Hall and Jaffe (2018).

stock of knowledge, which depreciates because of physical disappearance (e.g., death of a scientist in case of tacit knowledge, fire in case of codified knowledge) or because of obsolescence (as new knowledge replaces old knowledge). A large literature has considered this stock of knowledge as a determinant of productivity (for surveys of this literature, see, e.g., Griliches [1995], Hall [1996], and Hall et al. [2010]).

Besides serving as a measure of innovation input, R&D can also be considered as a way to assimilate knowledge so as to be better able to absorb outside knowledge. In this regard, it is like an investment in education to increase the absorption capacity. This dual aspect of R&D investment has been articulated by Cohen and Levinthal (1989).

It is not always crystal clear what is, and what is not, considered as R&D. In the (2015) version of the Frascati Manual, five conditions are stated to characterize R&D: It has to be aimed at new findings, it has to be based on original concepts and hypotheses, it has to be uncertain about the final outcome, it has to be planned and budgeted, and it has to lead to results that can be reproduced. For a long time, the inclusion or not of software in R&D was a matter of discussion. In the new version of the Manual, software is included if it satisfies the five criteria just mentioned. Another limitation of R&D is that more inputs are needed to innovate than just doing R&D. The Oslo Manual (the latest version of which is OECD 2018) has made a serious attempt in this direction by enlarging the scope of innovation expenditure.

The R&D surveys are, unlike the innovation surveys, supposed to cover all R&D performers in a country, past observed R&D performers as well as new suspected R&D performers because they have for instance applied for R&D tax credits, subsidies or other forms of government support for innovation. R&D statistics are regularly collected on a yearly basis. Small firms are underrepresented: First of all, R&D surveys are often limited to firms above a certain size in terms of number of employees; second, often a more concise questionnaire is sent to small firms; third, in some countries like Canada R&D from small firms are provided to the statistical offices by the tax department; and fourth, in other countries like The Netherlands the R&D statistics are collected in tandem with the innovation surveys in the years the innovation surveys take place—to avoid different numbers from two separate surveys—and they only cover so-called core R&D performers in the years between two innovation surveys. Moreover, the R&D statistics only cover formal R&D. Small firms without a formal R&D department might be doing some informal R&D and not bother reporting it to the statistical office.

2.2 Patent Statistics

In parallel to the literature on the returns to R&D, another branch of studies has explored the estimation of a knowledge production function, linking knowledge inputs in the form of R&D with knowledge outputs in the form of patents. Patents are used as a measure of knowledge output, which can then be inserted in the explanation of other economic variables like productivity or market value. The output measured here is closer to the notion of invention than to the notion of innovation. Patenting is a measure of protection of intellectual property. It may help in bringing new products or processes on the market, but it is not a requisite for it, nor is it sufficient to be successful in innovating. Moreover, patents may be applied for strategic reasons to create entry barriers (e.g., patent thickets), to be able to cross-license, or as signals of capability in order to attract outside funding.

Although some earlier studies had already tried to investigate the link between patents and productivity, the literature of patents as indicators of inventive performance really took off with the NBER work under the direction of Zvi Griliches (see in particular the 1984 NBER conference volume and his 1990 paper in the JEL).

Patents contain a lot of extra information besides the recording of a patent grant, the date and the technology class: applicant, assignee, inventor, number of claims, citations to previous patents and publications, priority application date, family information and many more (see Nagaoka et al. 2010). It is well known that the distribution of patent values is highly skewed. Therefore, it makes more sense to weigh the number of patents somehow, for instance, by giving more weight to patents that receive many forward citations.

When performing interindustry comparisons, one should be aware that in some fields, it is more difficult to patent and that some firms prefer not to patent. The 1987 Yale Survey on Industrial Research and Development (Levin et al. 1987) and the Carnegie-Mellon University R&D Survey of 1994 (Cohen et al. 2000) have clearly shown that patents are widely used in fields such as chemicals, drugs and computer and not so much in other fields, where firms prefer alternative means to appropriate the returns from investing in knowledge, such as being the first on the market or developing complex technologies. Similar results of patent concentration in a few sectors are reported by Arundel and Kabla (1998) for Europe. Applying for patents and especially defending one's patents against infringement can be costly and discourage many firms, especially small firms, from applying for patents.

Patent data have the advantage that they are easily available, for long periods of time, and that they contain lots of information on the content of the

patented invention, the timing of introduction, renewals and termination, the name and the location of the assignee and references to prior knowledge. All these pieces of information can be useful to infer the private and social value of a patent. The weakness of patent data is the selectivity of patenting, the difficulty of merging patent data with other firm-level data (technology classification versus industry classifications, disambiguation for matching on the basis of firm names).

2.3 Innovation Surveys[2]

The innovation surveys follow the guidelines of the Oslo Manual. They collect three types of information on innovation: innovation inputs, outputs and modalities.

The latest version of the Oslo Manual (OECD 2018) defines innovation as "a new or improved product or process (or combination thereof) that differs significantly from the unit's previous products or processes and that has been made available to potential users (product innovation) or brought into use by the unit (process innovation)." Product innovations encompass goods or services that have undergone significant improvements in one or the other functional characteristic such as quality, affordability, durability to name just a few. Process innovations refer to improvements in the business functions such as increased efficiency, meeting regulatory requirements or cost reductions. The Oslo Manual (OECD 2018) recognizes 6 types of business processes: production of goods and services, distribution and logistics, marketing and sales, information and communication systems, administration and management and product and business process developments. Organizational and marketing innovations, which were identified separately in the third version of the Oslo Manual, are now considered as part of process innovations. In contrast to patents, innovation measures the implementation and not just the invention of something new. Here also, there may be disagreements about what is included in this definition. Price changes due to external circumstances, seasonal and routine changes in the type of products sold, mere color changes or customization are not considered as innovations. Some scholars consider that any change in the way business is done is an innovation. There subsists thus a gray area in the definition of innovation.

[2]Since 2009 in the United States, the Business R&D and Innovation Survey, conducted jointly by the National Science Foundation/Science Resources Statistics (NSF/SRS) and the US Census Bureau, replaces the Survey of Industrial Research and Development by adding to the R&D survey some questions related to innovation. It is more an R&D survey than an innovation survey.

Innovation surveys collect data on innovation expenditure, which comprises besides the intramural and extramural R&D expenditure already collected in the R&D surveys, engineering, design and other creative activities, marketing and brand equity activities, IP-related activities, employee training activities, software developments and database activities, activities related to the acquisition of lease of tangible assets and innovation management activities (OECD 2018). Unfortunately, many of these items are not (yet) collected regularly by all firms and therefore difficult to quantify and very likely subject to substantial measurement errors. Think of employee training activities specifically for the production of new products or the use of new machines, not employee training activities in general.

The innovation surveys also collect information about the modalities of innovation, such as research collaborations, obstacles to innovation, sources of information, innovation objectives, presence of government support or environmental innovations.

Innovation surveys are supposed to be representative regarding size, industry and in some countries even regional distribution, based on stratified random sampling, above a certain minimal size threshold. They are conducted every two years now in EU countries (every four years previously) and on a more irregular basis in many other countries. A few countries have yearly data (Germany since 1993,[3] Spain since 1990,[4] United States since 2009).

The innovation survey data have certain characteristics that are important to keep in mind when using them in empirical research. First, they are to a large extent subjective data: The definition itself of what is an innovation leaves room for interpretation, whether a product is new to the firm or new to the market depends on the perception of what the relevant market is, and some data asked in the surveys are not systematically collected by firms (such as training for innovation or the share of sales due to new products) and therefore more guesstimates than hard data. The likely presence of errors in variables in the innovation survey data and the ensuing attenuation bias in the estimation of the relationship between innovation and productivity has been formally shown in Mairesse et al. (2005) and Mairesse and Robin (2017).

Second, contrary to the R&D and patent statistics, few of the data are quantitative to reveal something about the extent of the innovation success. Among the various types of innovation, there is a quantitative measure only

[3]The German Mannheim Innovation Panel is managed by the ZEW-Leibniz Center for European Economic Research.

[4]The Spanish ESSE (Encuesta sobre Estrategias Empresariales) Survey on Business Strategies has been conducted since 1990 by the Ministry of Industry and the SEPI Foundation.

for product innovation, the share of total sales due to new products. A few countries have quantitative measures for process innovation, namely the share of cost reduction due to new processes. For most countries, though, only dichotomous information exists for process innovation. While binary variables are less informative than continuous variables, it can, however, be argued that the errors in variables are less distorting with binary information.

Third, there is a timing problem, in the sense that innovation refers to a three-year period, whereas the few quantitative variables refer only to the last of the three years. It makes for instance little sense to explain the fact that a firm has innovated sometime over a three-year time span by the amount it spent on R&D in the last year of that period. Fourth, there is a potential selectivity issue as some variables are collected only for innovators. For example, no data on R&D are collected for firms that do not declare to have been innovative. Fifth, it is difficult to conduct panel data analysis with the innovation survey data because of the stratified random sampling. Only large firms (e.g., above 250 employees; the threshold depends on the country) will be approached in every wave. Smaller firms might randomly not be included in every wave. This systematic inclusion of larger firms may create a selection bias in the results obtained. Sixth, the structure of the questionnaire of the innovation surveys, the wording of the questions, the sampling and the mere mandatory nature of these surveys differ across countries more than the R&D surveys, rendering the innovation surveys less comparable internationally than the R&D surveys.

A general problem faced when relating innovation indicators to innovation or economic performance is the endogeneity of innovation. Some variables that drive innovation efforts also drive directly economic performance, and there may be a two-way relationship between the two variables. Many other variables contained in the innovation surveys may also be subject to endogeneity. Hence, unless the innovation survey data can be merged with other statistics or be made into a longitudinal dataset, there will be a problem of instrumenting the endogenous variables.

Contrary to patent data, R&D and innovation survey microdata are not as easily accessible for reasons of confidentiality. It is therefore difficult to merge innovation survey data from different countries to conduct international comparisons, unlike what can be done with business register data like ORBIS/AMADEUS from Bureau van Dijk, the Business Environment and Enterprise Performance Survey (BEEPS) database from the World Bank and the European Bank for Reconstruction and Development, or the EU Industrial R&D Investment Scoreboard database managed by the Joint Research Center of the European Commission.

Despite this long list of challenges that the user of innovation survey data should be aware of, these data contain new statistics, which have enlarged our understanding of the determinants and the effects of innovation on economic performance, as we shall see in Sect. 3.

2.4 Other Data

Literature-Based Innovation Counts

Another measure of innovation output is the literature-based innovation output (LIBO) indicator, which counts innovation announcements that are published in trade and technical journals (Coombs et al. 1996; Santarelli and Piergiovanni 1996). One of the first to introduce it were Kleinknecht et al. (1993). This indicator offers some advantage compared to the innovations surveys: It is less subjective than the innovation outputs from the innovation surveys since it is based on published material and verifiable, it gets recorded soon after the introduction on the market and not one or two years afterward, it can cover the small firms better than the innovation surveys (as shown by Kleinknecht 1987), and in principle, it could provide more details about the innovation itself. It has, however, the disadvantage that announcements are to some extent subject to self-selection, confined to product innovations, cover tangible goods more than intangible services, focus more on inputs and capital goods, are often biased toward major innovations and are not systematically collected and readily available for all countries.

Actually, a forerunner of the LIBO count data was the Science Policy Research Unit (SPRU) innovation database. This dataset was set up as follows. Experts from industry were asked to identify significant technical innovations that were commercialized in the UK between 1945 and 1983. Firms producing these innovations were then approached to provide information about the innovation and characteristics of the firm (Robson et al. 1988). This database ultimately led to the development of the innovation surveys, which are no longer based on specific innovations but on firms that innovate or not. In other words, the innovation surveys follow the *subject* approach, collecting information about a particular firm, instead of the *object* approach, where the basic statistical unit is an innovation.

Bibliometrics/Scientometrics

There is a branch of research called bibliometrics/scientometrics that uses publications and citations from databases such as Google Scholar, Scopus

or Web of Science to measure the quantity and the impact of scientific research. These indicators are used for monitoring scientific research output and for measuring productivity of scientific research in universities, research labs, individual researchers and scientific fields, or for measuring technology transfer or collaborations between research institutes and enterprises, more than for explaining the role of innovation in explaining productivity variations within and between firms. These indicators can, however, be helpful as indirect indicators of the connectivity between researchers or the quality of other indicators. To cite one example, Callaert et al. (2006) have looked at backward citations to non-patent references in patent applications to assess the science intensity of patents.

Inventor Surveys

The inventor surveys collect data on the inventors obtained from the patent databases, e.g., the PatVal Survey for six European countries (Giuri et al. 2007), the RIETI-Georgia Tech inventor survey (Walsh and Nagaoka 2009) for the United States and Japan. The aim of those surveys is primarily to gather information about inventors such as profiles, motivations, mobility, performance and perceived value of the inventions. Inventor survey data have been used as an alternative to patent citations for measuring the value of a patent, sources of knowledge and knowledge spillovers.

Market for Technology

Instead of conducting their own R&D, firms may decide to buy know-how instead on the market for technology. The innovation surveys contain some binary and continuous data on the purchase of patents and investments related to new technologies among the innovation expenditures. Licensing is another way to purchase outside technologies. No systematic data on licensing deals exist. The European and Japanese Patent Offices (EPO and JPO) organized a survey of licensing among patent holders in 2007 (Zuniga and Guellec 2009). Arqué-Castells and Spulber (2018) use data on patent trades from USPTO, licensing deals from the Securities and Exchange Commission (SEC) filings (ktMINE's licensing database) and cross-licensing data from the SEC forms, as well as Google searches, to construct connections in the market for technology. They find that when the returns on the markets for technology, which diffuse technological change, are internalized, the private and social rates of return on R&D increase substantially, by as much as 50% and 100%, respectively.

Technology Adoption and Diffusion

One way to foster technological change is to develop new products, services or technologies, and another one is to adopt existing technologies and ensure their diffusion throughout the economy. Surveys on the adoption of advanced technologies in manufacturing have been conducted in a number of countries. They do not identify transactions and amounts paid, but they identify whether a firm has used a range of advanced technologies. Empirical studies examining the link between the adoption of advanced technologies and productivity growth in manufacturing conclude that there is a positive link between the two variables (e.g., Baldwin and Sabourin (2002) for Canada, Bartelsman et al. (1998) for The Netherlands).

User Innovation

Firms are user innovators if they develop a process innovation for their own use or if they adopt a process and adapt it for their own use. A sizeable proportion of firms are user innovators, as high as 54% in high-tech Dutch small and medium enterprises (de Jong and von Hippel 2013). User innovators are more prone than commercial innovators to share their findings, and the adoption rate of user innovations is also higher than adoption rates in general.

3 Innovation and Productivity[5]

In this section, we shall examine what we have learned from R&D, patents, innovation surveys and innovation count data regarding the link between innovation and productivity.

3.1 Studies Based on R&D Data

The various indicators of innovation that have been listed above have been used in various ways to measure their impact on economic performance at the firm, sector or country level. In endogenous growth models, productivity

[5]Part of this section is based on Mohnen (2018), "The role of research and development in fostering economic performance. A survey of the macro-level literature and policy implications for Finland," Report submitted to OECD, February 2018.

growth is in part due to R&D efforts that are only undertaken if the costs of engaging in R&D (those can be variable, fixed or even sunk costs) do not exceed the returns from doing R&D. R&D generates innovation in the form of new intermediate inputs or new consumer goods, the variety of which increases productivity or consumer utility. In parallel to this love for variety approach, a Schumpeterian creative destruction approach has been developed in which new products replace old products because of superior quality instead of just increasing the range of products in the market and diminishing the margins made on old products (see Aghion and Howitt 1998; Barro and Sala-i-Martin 2004). There is also a debate in this literature between the contenders of the semi-endogenous and the fully endogenous R&D-based growth models, the former arguing that the returns to R&D are decreasing, the latter defending the assumption of constant returns to R&D. Ha and Howitt (2007) show evidence in favor of the Schumpeterian fully endogenous growth models, whereas Bloom et al. (2018) illustrate the declining productivity of R&D in a number of research fields. Nonetheless, so they argue, endogenous growth can survive because of the non-rival nature of knowledge.

Spillovers play an important role in R&D-based growth models. They can be positive as knowledge gets transmitted between agents or over generations or when rents occur because of imperfect price discrimination or network externalities. They can also be negative because of decreasing returns, duplication, obsolescence or market stealing. A number of macrostudies based on assumptions regarding these various forces have simulated the societal effects of R&D on economic growth. Depending on whether the positive or the negative externalities dominate, there is private underinvestment or overinvestment in R&D (see Montmartin and Massard 2015).

Even before these theoretical developments in the modeling of endogenous economic growth took place, empirical studies were devoted to estimate the returns to R&D starting with Griliches (1964) and Mansfield (1965). The underlying model is an extended production function with as additional input the stock of knowledge obtained from R&D expenditure. The stock of knowledge depreciates when tacit knowledge gets lost for instance with the death of a scientist[6] or when through obsolescence new knowledge supersedes old knowledge. The idea is to estimate the

[6]Recent work on team capital confirms this loss of tacit knowledge. Azoulay et al. (2010) find that the premature death of a superstar scientist reduces by 5–8% the quality-adjusted publication record of his (her) collaborators. In the same vein, Jaravel et al. (2018) find that the unexpected death of an inventor decreases the co-inventors' earnings by 4% and their citation-weighted patents by 15% after 8 years.

increment in production or value added due to a marginal increase in the stock of knowledge. If this marginal productivity remains constant over time, it can also be interpreted as the internal rate of return that equates costs and revenues gross of the depreciation rate of the stock of knowledge. When multiplying this marginal productivity by the R&D over output ratio, one gets the elasticity of output with respect to R&D, which multiplied by the growth rate of the R&D stock measures the contribution of R&D to output or TFP growth in growth accounting.

Spillovers are captured by including as an additional argument in the production function the R&D stock accumulated outside of the firm. This is usually done by constructing a weighted average of the R&D stocks of other R&D performers (plants, firms, sectors, regions or countries depending on the level of aggregation), unless one wants to estimate separate spillover sources, which can quickly become difficult to identify as the allowed number of sources increases. Various weighting schemes have been experimented with depending on the assumed channel of transmission of the spillovers: geographical proximities, R&D collaborations, co-patenting, correlations of positions in the patent classes or in the lines of business, patent citations, interindustry transactions, international trade, foreign direct investment, to name the most popular ones. If the outgoing R&D externalities are added to the private rate of return to R&D, one obtains a social rate of return to R&D, that is, the return to society at large.

The rate of return to R&D has been estimated in a variety of ways. We briefly list below several of the major differences in specification and the possible effects they could have on the estimated returns to R&D. For a more thorough and detailed discussion of these issues, the reader is referred to the initial presentation of the whole framework in Griliches (1979) and to the survey by Hall et al. (2010).

Regarding the specification, most studies have used a Cobb-Douglas production function. Some have used a translog or other second-order approximations of a general production function, which allow for complementarities or substitutions between R&D and other inputs. Some studies have preferred assuming a constant elasticity of output with respect to productivity rather than a constant marginal productivity of R&D. Estimates seem to be more stable with a constant elasticity specification, implicitly assuming a declining marginal productivity of R&D. Some studies have favored a dual representation of technology, conditional on variable factor prices and, maybe more contentiously, on the exogenous level of production in lieu of the input levels. A system of demand equations can then be estimated, which increases the number of degrees of freedom. Sometimes a

mixture of variable and quasi-fixed inputs is allowed for. A few studies have opted for an intertemporal model of decision-making to derive the optimal path of knowledge accumulation, which yields the specification of the demand for R&D equation.

Regarding the data, the earlier studies used sector or aggregate country data. Nowadays, the majority of studies are based on firm data or even on establishment data. At a higher level of aggregation, one would expect higher rates of return because of internalized spillovers, but this is not systematically the case. Ideally, the traditional inputs should be cleared of their R&D component to avoid R&D double-counting (Schankerman 1981). This is rather rarely done at the cost of yielding underestimates of the returns to R&D. A crucial element in the estimation of the rate of return to R&D is the assumed depreciation rate. At the beginning of this literature, when time series on R&D were still relatively short, a zero rate of depreciation was often assumed to obviate the need to construct a stock of knowledge. Later, studies constructed R&D stocks assuming constant—over time and space—R&D depreciation rates. The latest studies obtain time- and industry-specific R&D depreciation rates (Li and Hall 2017).

The production function or the dual representation of technology has been estimated in levels or in growth rates. Estimates are generally higher, more stable and more likely to be significant when based on levels rather than growth rates. Most studies are based on time series data, exploiting only the temporal variation, some use only cross-sectional data, the more recent studies exploit panel data, where both temporal and cross-sectional variations can be exploited and individual effects can be controlled for. Typically, lower returns are obtained in the within than in the between variation. Some studies have controlled for other factors that may affect productivity, such as human capital, organizational capital, ICT equipment, R&D spillovers or sector specificities. The returns to R&D tend to drop when these other variables are introduced.

Over the last 50 years, many empirical papers have been devoted to the estimation of the private and the social rates of return to R&D (see the survey by Hall et al. [2010] and the meta-analyses by Wieser [2005], Koopmans and Donselaar [2015], and Ugur et al. [2016]). Despite the large heterogeneity in the results obtained, the following seem to be reasonable orders of magnitude. The private rate of return on R&D exceeds the normal rate of return and is in the 10–30% range. Estimates of the elasticity of output with respect to R&D are largely consistent with those of the rates of return and hover around 0.10. Given these estimates and the growth in R&D stock, the contribution of R&D to TFP growth is expected to be

in the range of 10–15%. The social rate of return exceeds the private rate of return by a factor of 50–100%. Rates of return are found to be heterogeneous. They are generally found to be higher for private than for public R&D and for basic R&D than for applied R&D or development. The estimated elasticities are generally higher in high-tech, i.e., R&D-intensive, than in low-tech sectors (e.g., Ortega-Argilés et al. 2015), but according to the results reported by Wieser (2005) and Ugur et al. (2016), the associated rates of return are not necessarily different between the two sectors. Rates of return may differ across countries because of differences in distance to the frontier (Griffith et al. 2004), industrial structure or national innovation systems (Kokko et al. 2015). Countries may benefit from international R&D spillovers. As shown in Mancusi (2008), laggard countries are mainly the beneficiaries, depending on their absorptive capacity, whereas technological leaders are mainly the source of international R&D spillovers.

The 2008 revision of the National Income and Product Accounts treats R&D as an investment and no longer as an expenditure. Fraumeni and Okubo (2005) have focused on the contribution of R&D in the new national income accounting. For the United States over the period 1961–2000, they arrive, on the expenditure side, at a contribution of R&D investment to corrected GDP between 2% and 7% depending on the scenarios and, on the income side, at a contribution of the returns on R&D to corrected GDP between 4% and 15%. Corrado et al. (2013) follow the approach of Corrado et al. (2009) and consider three types of intangible assets: (i) computerized information (software, databases), (ii) innovative property (research and development, mineral exploitation, copyright and license costs and other product development, design and research expenses) and (iii) economic competences (brand equity, firm-specific human capital and organizational structure). They have capitalized the investments in these intangibles under some assumptions regarding deflators and depreciation rates. They find that innovative property (including R&D) accounts for a proportion of labor productivity growth that ranges from 4.5% in the UK to 12.5% in the United States.

Besides the extended production function approach, there are at least three other approaches that are worth signaling: the stochastic efficiency frontier, the market value and the stochastic productivity residual. The stochastic efficiency frontier estimates both the outward shift of the frontier and changes of positions with respect to the frontier. Kumbhakar et al. (2012) estimate a parametric stochastic efficiency frontier instead of a production function. For a sample of top European R&D investors between 2000 and 2005, they show that in high-tech sectors, R&D mainly shifts

out the frontier, whereas in low-tech sectors its role is mainly to bring firms closer to the frontier. Many studies have also looked into whether valuations of firms in the stock market are related to the volume of their R&D capital stocks in publicly traded firms. The underlying model is a market value equation that depends on the replacement value times Tobin's q, which depends on knowledge capital (see Griliches 1990; Hall 2000). Although this method can only be applied for publicly traded firms, it has the advantage of including expected future returns. Positive effects of R&D have been estimated for many countries, although these estimated coefficients are lower than one, suggesting overinvestment, insufficient shareholder protection or too low R&D depreciation rates used in the construction of the R&D capital stocks (see Hall and Oriani 2006). The last approach that we want to mention models R&D no longer as a capital stock, which affects productivity in a linear and deterministic fashion, but as an investment that affects the distribution of TFP. Using this kind of framework, Doraszelski and Jaumandreu (2013) find that in most Spanish industries the return to R&D is higher, the higher is past productivity and that the mean expected productivity is higher for R&D-performing than for non-R&D-performing firms. The net rate of return to R&D varies across industries but averages around 40%, being higher in industries where the uncertainty is higher.

3.2 Studies Based on Patent Data[7]

In principle, the methods used in the previous section could be applied to the stock of patents as a measure of the stock of knowledge in lieu of the R&D stock. In this way, the patent stock could be related to productivity, market value, movements to the efficiency frontier or the Markov process governing the stochastic productivity residual. The fact that the distribution of patent values is highly skewed, with very few patents being worth a lot, militates in favor of using R&D instead of patents to explain TFP or market value, because the errors in variable problem are higher for patents than for R&D. Hall et al. (2005) have compared the effect of R&D, patents and citations on the market value of firms and found that a percentage point

[7]Patent data have been used for other topics than their link to R&D and productivity, like the strategic use of patents (pre-emptive patenting, patent trolls, patent litigations, patent thickets), or policies for protecting intellectual property (patent length, patent breath, patentability); see Hall and Harhoff (2012). We shall limit ourselves to the use of patents as indicators of innovation and their link to variations in productivity.

increase in the R&D/assets ratio leads to a 0.8% increase in market value, that an extra patent per million $ of R&D boosts market value by about 2%, and an extra citation per patent boosts it by over 3%. They also find that the market values are particularly correlated with citations that cannot be predicted from past citations. Although patent counts are not as good predictors of market value as R&D, they nevertheless add to the understanding of market values.

What has also been examined is the link between patents and R&D, one version of the so-called knowledge production function (Griliches 1990). It has been found that patents are correlated with R&D and that there is hardly any lag between the two. Here again, the relationship is less visible in the within temporal dimension. In the cross-sectional dimension, the relationship between patents and R&D is higher for small than for large firms, because of selectivity (observing the best small firms) and more frequent use of informal IP protection in large firms and informal R&D in small firms.

Patents can be very useful for estimating R&D spillovers. There are two ways in which this can be done. The first is to measure a spatial correlation of firms in the patent space, i.e., the vector positions of firms in patent classes. This idea goes back to Jaffe (1986). The idea is that the more firms patent in the same or in close patent classes, the more they perform similar research and benefit from each other's research. The second way patents can be used in connection to R&D spillovers is by way of patent citations. Citations to previous patents can be considered as proxies for knowledge flows between firms. This approach had been used to estimate spillovers across industries (Scherer 1982), countries (Jaffe et al. 1993; Verspagen 1997) or regions (Peri 2005). Patent citations tend to be localized, and therefore, if they are supposed to reflect knowledge flows, they point to geographical spillovers that decrease with distance to the origin. Peri (2005) finds that only 20% of the knowledge generated in a region flows out of it even though knowledge flows are much less localized than trade flows.

Using the Google Patent database, Kogan et al. (2017) infer the value of patents from the stock market reactions three days after patents are issued. A firm's innovation is measured as the sum of the values of all the patents granted to a firm normalized by its size. The authors find that a one standard deviation increase in a firm's innovation is associated with a 2.4% increase in a firm's revenue-based productivity, whereas a one standard deviation increase in innovation by a firm's competitors is followed by a 1.7% drop in productivity over five years. At the macrolevel, they find that a one standard deviation increase in macroinnovation leads to a 3.4% increase in TFP growth in the next 5 years.

3.3 Studies Based on Innovation Survey Data

With the advent of the innovation surveys, which started to be collected in many countries in the early 1990s, it became possible to relate productivity with measures of implemented innovation output instead of just innovation inputs. Actually, the production function relating productivity to innovation output could be combined with a knowledge production function relating innovation input (R&D or innovation expenditure) with innovation output. This structural model was first proposed by Pakes and Griliches (1984) using patents as innovation outputs and later implemented by Crépon et al. (1998), using patents and the share of innovative sales as alternative measures of innovation output, in what has come to be known as the CDM model. It treats the endogeneity of R&D and innovation output by having an equation explaining the amount of R&D, one that explains the intensity of innovation and one that explains productivity in growth rates or in levels. Moreover, some firms happen to do no R&D and many are not innovative. This selectivity issue is also handled in the CDM model using tobit models or Heckman's two-step approach. The CDM framework allows for the use of binary and continuous data for innovation inputs and or outputs and in principle for multiple sources of innovation.

The original CDM model is a recursive model without feedback from productivity to R&D or innovation. It may well be that productive firms are more innovative because they can afford to finance innovation projects. Several attempts have been made to let this happen by introducing past productivity in the innovation input or output equations (Baum et al. 2017; Raymond et al. 2015; Cainelli et al. 2006). Another generalization of the CDM model consists in allowing for lags in the relationships among R&D, innovation and productivity, as well as for persistence in innovation and productivity. It is important in that case to allow for unobserved heterogeneity so as to avoid spurious persistence. Persistence seems to be correlated with the intensity of innovation as it is found to be more pronounced for R&D-performing innovative firms (Peters 2009), in high-tech industries (Raymond et al. 2010) and for radical innovators (Zhen 2018).

When continuous measures of innovation output are used, the typical orders of magnitude of the elasticities of output with respect to innovation are between 0.10 and 0.25, indicating that a 10% increase in innovation output (sales of new products per employee) increases labor productivity by 1–2.5% (Mohnen and Hall 2013). The elasticity of productivity with respect to innovation output declines when other factors like capital stock or human capital are controlled for. As was also mentioned for R&D, lower elasticities

are found when the regression is in growth rates rather than levels of productivity. The innovation survey allows for various levels of novelty of product innovations by distinguishing products new to the firm and products new to the market. With continuous data, no major differences are found regarding the level of novelty. When only binary data on innovation output are available, innovation generally increases productivity significantly, whatever kind of innovation is considered. Peters et al. (2017) report that in German high-tech industries, it is product innovation that increases productivity, and in low-tech industries, it is process innovation. As Jaumandreu and Mairesse (2017) actually argue and show, it is difficult to identify separately the effect of different types of innovation, partly because we know too little to instrument each type of innovation output by different exogenous variables and partly because different types of innovation are often introduced simultaneously.

On French data, Mairesse et al. (2005) have shown that the rates of return to R&D calculated from the CDM model are consistent with those obtained from the reduced form model where R&D enters the production function directly. What these innovation surveys have also revealed is that especially for low and medium technology firms, small- and medium-sized firms and firms in developing countries, non-R&D is an important input in the innovation process besides formal R&D. Instead of relying on their own R&D, these firms buy outside technologies and invest in advanced manufacturing technologies, licensing and training to advance their state of knowledge (Santamaría et al. 2009; Huang et al. 2010). The CDM model has recently been generalized by Peters et al. (2017) in the direction of making the effect of R&D on innovation and of innovation on productivity stochastic. Their model allows for firms to be innovative without doing R&D; as a matter of fact, on German data they find that this is the case for 22% of the firms. Firms that do R&D are more likely to be innovative, but R&D is not a sufficient condition for being innovative. The probability of turning out not to be innovative is 10% in low-tech industries and 20% in high-tech industries. The long-run rate of return to R&D is calculated as the relative difference in the expected firm value between firms that do and those that do not do any R&D. In the high-tech industries, the median rate of return to R&D is 6.7%. In low-tech industries, the corresponding figure is 2.8%. They also find a lot of heterogeneity between firms and thereby rejoin Baum et al. (2017), who report that the relationship between innovation and productivity differs across industries. The international comparison study performed on 18 OECD countries also found heterogeneity across countries, types of sectors and sizes of firms with generally larger effects of innovation on productivity in manufacturing than in services (OECD 2009). The positive links between innovation input, innovation output and productivity are also obtained on Latin American data, but the semi-elasticity of productivity with respect to dichotomous

measures of innovation tends to be higher in Latin American than in European countries, reflecting a greater productivity gap that could be overcome by innovation in the former countries (Crespi and Zuniga 2012).

There is mixed evidence regarding the existence of any complementarity between different types of innovation, meaning that the return from one type would increase in the presence of the other type. Ballot et al. (2015) find some complementarity between product and process innovation in France and in the UK, but only complementarity between product (not process) and organizational innovation in France (not the UK). Peters et al. (2017) find no sign that the simultaneous introduction of product and process innovation has any additional effect in German firms, whereas Schmidt and Rammer (2007) conclude that product and process innovations lead to higher cost reductions or more novel (new-to-market) product innovations when combined with both organizational and marketing innovations.

3.4 Innovation Count Data

One of the first studies using counts of new products is by Comanor and Scherer (1969). They used two measures of new product counts corresponding to the notions of new to the market and new to the firm: the number of new chemical entities introduced by each pharmaceutical firm from 1955 to 1960, with each new product weighted by its sales during the first two calendar years following introduction, and a similar broader measure that includes combinations of active ingredients, new dosage forms and products that merely duplicate those already introduced by competing firms as well as new chemical entities. They found significant positive correlations between the three measures even after controlling for firm size.

Acs and Audretsch (1988) exploit count data on announced innovations compiled by the US Small Business Administration from listings in hundreds of trade journal. They report a higher correlation between innovation counts and patents than between innovation counts and R&D. When controlling for other determinants, they obtain an elasticity of innovation counts with respect to corporate R&D close to 0.5. Using the SPRU innovation count database, Geroski (1991) and Sterlachinni (1989) find a positive correlation between the number of innovations used in an industry and its productivity growth.[8]

[8]Sjöö (2016) examines whether there was an industrial renewal in Sweden between 1970 and 2007 in terms of degree of novelty, volume, firm size, concentration and industrial origin on the basis of some 4000 innovations introduced in Sweden during this time period. She does not relate innovations to productivity growth.

4 Conclusion and Discussion

Whatever the innovation indicator, there will always be the problem that part of the variation of productivity reflects mismeasured prices. Few micro-datasets contain product prices. To the extent that industry deflators incorrectly measure firm-specific price changes on the input or on the output side productivity gets over- or underestimated. This problem is magnified when it comes to innovation. First, prices of new products are hard to measure, second, quality changes are difficult to dissociate from pure price changes, and third, part of revenue productivity growth can be due to market power instead of efficiency in the production of goods or services.

Over the last 50 years, efforts have been made to collect indicators of innovation inputs and outputs in a systematic and standardized way. R&D surveys are conducted in almost all countries, and innovation surveys are conducted on a regular basis in more and more countries. Patent applications have soared, thereby collecting useful data on inventions in technology classes, citations and patent renewal fees to infer the value of patents and to measure knowledge spillovers. With progress in digital technology, information on patents and other IP tools like trademarks, licenses and utility models can be easily stored and made available worldwide. In the future, big data will allow the examination of innovation from other angles, such as consumption patterns and networks.

The choice between indicators depends on the purpose of their use. In this paper, we confined ourselves to explaining TFP growth. They could also be used to assess domestic and international competitiveness, employment, standard of living, development or inequality in the distribution of income. Policy makers tend to concentrate on a particular indicator for monitoring and benchmarking innovative capabilities, for instance, the R&D over GDP ratio. This rather narrow view of technological capabilities neglects at least three facts: First, some industries are more R&D-intensive than others, and a country might be specialized in low R&D-intensive industries; second, what matters is not just R&D generation but rather R&D use and a country may decide to buy knowledge in the technology market rather than doing R&D itself; and third, many digitally based innovations are services, which do not require much R&D but developments of connectivity, multi-sided markets and integration of technologies.

Even when it comes to explaining TFP growth, there is not one best indicator. As we have seen, every indicator has its specificities, strengths and weaknesses. Some measure the inputs, and others measure the outputs of technological of innovation; some are easily available, and others require

special permissions; some are collected regularly and others only occasionally; some present themselves as panel data and others only as cross-sections. They may be biased toward large firms or publicly listed firms. They may pertain to a particular date or to a longer period. They may reflect a verifiable transaction or they may represent guestimates. And the list goes on. One solution is to construct an index based on these various indicators. While it may do a good job in terms of monitoring and benchmarking, it does not exploit the full information contained in multiple indicators, which would lead to a better understanding of the links between them and the ultimate performance measure one seeks to explain.

Improvements will be made in the future thanks to the ease of sharing and storing information. New indicators will be developed such as the tracing of the value chains for many goods, data on functionalities rather than services and integration of worldwide operations of multinational firms. As much as possible, we should try to strive for longitudinal data that can be merged with other data.

The present state of knowledge confirms Schumpeter's and long before him John Rae's vision of innovation as driver of economic growth. Whatever innovation indicator we select, the evidence overwhelmingly shows that in the long run, innovation is correlated with TFP growth whether at the firm or at the aggregate level.

Acknowledgements This paper was written under EU-funded project GROWINPRO, GA 822781. Part of this paper was written while I was visiting Renmin University of China, which I would like to thank for its hospitality. This chapter is not meant to be an exhaustive survey of the literature. Scholars should not feel offended if some of their work on these topics is not mentioned. References serve mainly to illustrate certain points. I thank Thijs ten Raa for encouraging me to write this paper. I am grateful to a referee for his/her constructive comments on a first version of this paper.

References

Acs, Z., and D. Audretsch. 1988. Innovation in large and small firms: An empirical analysis. *American Economic Review* 78 (4): 678–690.

Aghion, P., and P. Howitt. 1998. *Endogenous growth theory*. Cambridge: MIT Press.

Arque-Castells, P, and D. Spulber. 2018. Widgets and wodgets: Technology markets and R&D spillovers. Northwestern Law & Econ Research Paper No. 18–18.

Arundel, A., and I. Kabla. 1998. What percentage of innovations are patented? Empirical estimates for European firms. *Research Policy* 27 (2): 127–141.

Azoulay, P., J.S. Graff Zivin, and J. Wang. 2010. Superstar extinction. *Quarterly Journal of Economics* 125 (2): 549–589.

Baldwin, J., and D. Sabourin. 2002. Advanced technology use and firm performance in Canadian manufacturing in the 1990s. *Industrial and Corporate Change* 11 (4): 761–789.

Ballot, G., F. Fakhfakh, F. Galia, and A. Salter. 2015. The fateful triangle: Complementarities in performance between product, process and organizational innovation in France and in the UK. *Research Policy* 44 (1): 217–232.

Barro, R.J., and X. Sala-i-Martin. 2004. *Economic growth*, 2nd ed. Cambridge: MIT Press.

Bartelsman, E., G. van Leeuwen, and H. Nieuwenhuijsen. 1998. Adoption of advanced manufacturing technology and firm performance in The Netherlands. *Economics of Innovation and New Technology* 6: 291–312.

Baum, C.F., H. Lööf, P. Nabavi, and A. Stephan. 2017. A new approach to estimation of the R&D-innovation-productivity relationship. *Economics and Innovation and New Technology* 26 (1–2): 121–133.

Bloom, N., C. Jones, J. Van Reenen, and M. Webb. 2018. Are ideas getting harder to find? NBER Working Paper 23782.

Cainelli, G., R. Evangelista, and M. Savona. 2006. Innovation and economic performance in services: A firm-level analysis. *Cambridge Journal of Economics* 30 (3): 435–458.

Callaert, J., B. van Looy, A. Verbeek, K. Debackere, and B. Thijs. 2006. Traces of prior art: An analysis of non-patent references found in patent documents. *Scientometrics* 69 (1): 3–20.

Cohen, W.M., R.R. Nelson, and J. Walsh. 2000. Protecting their intellectual assets: Appropriability conditions and why U.S. manufacturing firms patent (or not). NBER Working Paper 7552.

Cohen, W.M., and D.A. Levinthal. 1989. Innovation and learning: The two faces of R&D—Implications for the analysis of R&D investment. *Economic Journal* 99: 569–596.

Comanor, W., and F.M. Scherer. 1969. Patent statistics as a measure of technical change. *Journal of Political Economy* 77 (3): 329–398.

Coombs, R., P. Narandren, and A. Richards. 1996. A literature-based innovation output indicator. *Research Policy* 25 (3): 403–413.

Corrado, C., J. Haskel, C. Jona-Lasinio, and M. Iommi. 2013. Innovation and intangible investment in Europe, Japan, and the United States. *Oxford Review of Economic Policy* 29: 261–286.

Corrado, C.A., C.R. Hulten, and D.E. Sichel. 2009. Intangible capital and U.S. economic growth. *Review of Income and Wealth* 5 (3): 661–685.

Crépon, B., E. Duguet, and J. Mairesse. 1998. Research, innovation and productivity: An econometric analysis at the firm level. *Economics of Innovation and New Technology* 7: 115–158.

Crespi, G., and P. Zuniga. 2012. Innovation and productivity: Evidence from six Latin American countries. *World Development* 40 (2): 273–290.

de Jong, J., and E. von Hippel. 2013. User innovation: Business and consumers. In *Handbook of innovation indicators and measurement*, ed. F. Gault. Cheltenham: Edward Elgar Publishing.

Doraszelski, U., and J. Jaumandreu. 2013. R&D and productivity: Estimating endogenous productivity. *Review of Economic Studies* 80 (4): 1338–1383.

Fraumeni, B., and S. Okubo. 2005. R&D in the national income and product accounts: A first look at its effect on GDP. In *Measuring capital in the new economy*, ed. C. Corrado, J.C. Haltiwanger, and D.E. Sichel, 275–322. Chicago: University of Chicago Press.

Gault, F. 2010. *Innovation strategies for a global economy: Development, implementation, measurement and management*. Cheltenham: Edward Elgar Publishing and International Development Research Center.

Gault, F. (ed.). 2013. *Handbook of innovation indicators and measurement*. Cheltenham: Edward Elgar Publishing.

Geroski, P. 1991. Innovation and the sectoral sources of UK productivity growth. *Economic Journal* 98: 375–390.

Giuri, P., M. Mariani, S. Brusoni, G. Crespi, D. Francoz, A. Gambardella, W. Garcia-Fontes, A. Geuna, R. Gonzales, D. Harhoff, K. Hoisl, C. Le Bas, A. Luzzi, L. Magazzini, L. Neste, O. Normaler, N. Palomeras, P. Patel, and B. Verspagen. 2007. Inventors and invention processes in Europe: Results from the PatVal-EU survey. *Research Policy* 36 (8): 1107–1127.

Griffith, R., S. Redding, and J. Van Reenen. 2004. Mapping the two faces of R&D: Productivity growth in a panel of OECD countries. *Review of Economics and Statistics* 86 (4): 883–895.

Griliches, Z. 1964. Research expenditures, education and the aggregate agricultural production function. *American Economic Review* 54 (6): 961–974.

Griliches, Z. 1979. Issues in assessing the contribution of research and development to productivity growth. *Bell Journal of Economics* 10 (1): 92–116.

Griliches, Z. 1990. Patent statistics as economic indicators: A survey. *Journal of Economic Literature* 28 (4): 1661–1707.

Griliches, Z. 1995. R&D and productivity: Econometric results and measurement issues. In *Handbook of the economics of innovation and technical change*, ed. P. Stoneman. Oxford: Blackwell Handbooks in Economics.

Ha, J., and P. Howitt. 2007. Accounting for trends in productivity and R&D: A Schumpeterian critique of semi-endogenous growth theory. *Journal of Money, Credit and Banking* 39 (4): 733–774.

Hagedoorn, J., and M.M.A.H. Cloodt. 2003. Measuring innovative performance: Is there an advantage in using multiple indicators? *Research Policy* 32: 1365–1379.

Hall, B. 2000. Innovation and market value. In *Productivity, innovation and economic performance*, ed. R. Barell, G. Mason, and M. O'Mahoney. Cambridge: Cambridge University Press.

Hall, B.H., A. Jaffe, and M. Trajtenberg. 2005. Market value and patent citations. *Rand Journal of Economics* 36 (1): 16–38.

Hall, B.H. 1996. The private and social returns to Research and Development. In *Technology, R&D and the economy*, ed. B.L.R. Smith and C.E. Barfield. Washington DC: The Brookings Institution.

Hall, B.H., and A.B. Jaffe. 2018. Measuring science, technology, and innovation: A review. *Annals of Science and Technology Policy* 2 (1): 1–74.

Hall, B.H., and D. Harhoff. 2012. Recent research on the economics of patents. *Annual Review of Economics* 4: 541–565.

Hall, B.H., and R. Oriani. 2006. Does the market value R&D investment by European firms? Evidence from a panel of manufacturing firms in France, Germany and Italy. *International Journal of Industrial Organization* 24 (5): 971–993.

Hall, B.H., J. Mairesse, and P. Mohnen. 2010. Measuring the returns to R&D. In *Handbook of the economics of innovation*, ed. B.H. Hall and N. Rosenberg, 1034–1082. Amsterdam: Elsevier.

Huang, C., A. Arundel, and H. Hollanders. 2010. How firms innovate: R&D, non-R&D and technology adoption. UNU-MERIT Working Paper 2010-027.

Jaffe, A. 1986. Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. *American Economic Review* 76 (5): 984–1001.

Jaffe, A.B., M. Trajtenberg, and R. Henderson. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108: 577–598.

Jaravel, X., N. Petkova, and A. Bell. 2018. Team-specific capital and innovation. *American Economic Review* 108 (4–5): 1034–1073.

Jaumandreu, J., and J. Mairesse. 2017. Disentangling the effects of product and process innovation on cost and demand. *Economics of Innovation and New Technology* 26 (1–2): 150–167.

Kleinknecht, A. 1987. Measuring R&D in small firms: How much are we missing? *Journal of Industrial Economics* 36 (2): 253–256.

Kleinknecht, A., J.O.N. Reijnen, and W. Smits. 1993. Collecting literature-based innovation output indicators. The experience in The Netherlands. In *New concepts in innovation output measurement*, ed. A. Kleinknecht and D. Bain, 42–84. London: Macmillan.

Kleinknecht, A., K. Van Montfort, and E. Brouwer. 2002. The non-trivial choice between innovation indicators. *Economics of Innovation and New Technology* 11 (2): 109–121.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman. 2017. Technological innovation, resource allocation and growth. *Quarterly Journal of Economics* 132 (2): 665–712.

Kokko, A., P. Gustavsson Tingvall, and J. Videnord. 2015. The growth effects of R&D spending in the EU: A meta-analysis. Economics Discussion Paper No. 2015-29, Kiel Institute for the World Economy.

Koopmans, C.C., and P. Donselaar. 2015. Een meta-analyse van het effect van R&D op productiviteit. *Economisch Statistische Berichten* 100 (4717): 518–521.

Kumbhakar, S., R. Ortega-Argilés, L. Potters, M. Vivarelli, and P. Voigt. 2012. Corporate R&D and firm efficiency: Evidence from Europe's top R&D investors. *Journal of Productivity Analysis* 37 (2): 125–140.

Levin, R.C., A.K. Klevorick, R.R. Nelson, and S.G. Winter. 1987. Appropriating the returns from industrial research and development. *Brookings Papers on Economic Activity* 3: 242–279.

Li, W.C.Y., and B. Hall. 2017. Depreciation of business R&D capital. *Review of Income and Wealth* (forthcoming). NBER Working Paper 22473.

Mairesse, J., and P. Mohnen. 2010. Using innovation surveys for econometric analysis. In *Handbook of the economics of innovation*, ed. B.H. Hall and N. Rosenberg, 1130–1155. Amsterdam: Elsevier.

Mairesse, J., P. Mohnen, and E. Kremp. 2005. The importance of R&D and innovation for productivity: A reexamination in light of the 2000 French innovation survey. *Annales D'Économie et de Statistique* 79 (80): 487–527.

Mairesse, J., and S. Robin. 2017. Assessing measurement errors in the CDM research–innovation–productivity relationships. *Economics of Innovation and New Technology* 26 (1–2): 93–107.

Mancusi, M.L. 2008. International spillovers and absorptive capacity: A cross-country, cross-sector analysis based on patents and citations. *Journal of International Economics* 76: 155–165.

Mansfield, E. 1965. Rates of return from industrial Research and Development. *American Economic Review* 55: 310–322.

Mohnen, P., and B. Hall. 2013. Innovation and productivity: An update. *Eurasian Business Review* 3 (1): 47–65.

Montmartin, B., and N. Massard. 2015. Is financial support for private R&D always justified? A discussion based on the literature on growth. *Journal of Economic Surveys* 29 (3): 479–505.

Nagaoka, S., K. Motohashi, and A. Goto. 2010. Patent statistics as an innovation indicator. In *Handbook of the economics of innovation*, ed. B.H. Hall and N. Rosenberg, 1084–1127. Amsterdam: Elsevier.

OECD. 1992, 1996, 2005 2018. *Oslo manual*, Paris, 1st, 2nd, 3rd, 4th Edition.

OECD. 2009. *Innovation in firms: A microeconomic perspective*. Paris: OECD.

OECD. 2015. *Frascati manual 2015: Guidelines for collecting and reporting data on Research and Experimental Development*. Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264239012-en.

Ortega-Argilés, R., M. Piga, and M. Vivarelli. 2015. The productivity impact of R&D investment: Are high-tech sectors still ahead? *Economics of Innovation and New Technology* 24 (3): 204–222.

Pakes, A., and Z. Griliches. 1984. Patents and R&D at the firm level: A first look. In *R&D, patents and productivity*, ed. Z. Griliches, Chicago University Press in the National Bureau of Economic Research Conference Series.

Peri, G. 2005. Determinants of knowledge flows and their effect on innovation. *Review of Economics and Statistics* 87 (2): 308–322.

Peters, B. 2009. Persistence of innovation: Stylized facts and panel data evidence. *Journal of Technology Transfer* 34: 226–243.

Peters, B., M. Roberts, V.A. Vuong, and H. Fryges. 2017. Estimating dynamic R&D demand: An analysis of costs and long-run benefits. *RAND Journal of Economics* 48 (2): 409–437.

Raymond, W., J. Mairesse, P. Mohnen, and F. Palm. 2015. Dynamic models of R&D, innovation and productivity: Panel data evidence for Dutch and French manufacturing. *European Economic Review* 78: 285–306.

Raymond, W., P. Mohnen, F. Palm, and S. Schim van der Loeff. 2010. Persistence of innovation in Dutch manufacturing: Is it spurious? *Review of Economics and Statistics* 92 (3): 495–504.

Robson, M., J. Townsend, and K. Pavitt. 1988. Sectoral patterns of production and use of innovations in the UK: 1945–1983. *Research Policy* 17: 1–14.

Santarelli, E., and R. Piergiovanni. 1996. Analyzing literature-based innovation output indicators: The Italian experience. *Research Policy* 25: 689–711.

Santamaría, J., M.J. Nieto, and A. Barge-Gil. 2009. Beyond formal R&D: Taking advantage of other sources of innovation in low- and medium-technology industries. *Research Policy* 38: 507–517.

Schankerman, M. 1981. The effects of double-counting and expensing on the measured returns to R&D. *Review of Economics and Statistics* 63: 454–458.

Scherer, F.M. 1982. Interindustry technology flows and productivity growth. *Review of Economics and Statistics* 64: 627–634.

Schmidt, T., and C. Rammer. 2007. Non-technological and technological innovation: Straneg bedfelllows? ZEW Discussion Papers 07-052. ZEW - Leibniz Center for European Economic Research.

Sjöö, K. 2016. Innovation and industrial renewal in Sweden, 1970–2007. *Scandinavian Economic History Review* 64 (3): 258–277.

Sterlachinni, A. 1989. R&D, innovations and total factor productivity growth in British manufacturing. *Applied Economics* 21: 1549–1562.

Ugur, M., E. Trushin, E. Solomon, and F. Guidi. 2016. R&D and productivity in OECD firms and industries: A hierarchical meta-regression analysis. *Research Policy* 45: 2069–2086.

Verspagen, B. 1997. Estimating international technology spillovers using technology flow matrices. *Review of World Economics* 133 (2): 226–248.

Walsh, J., and S. Nagaoka. 2009. Who invents?: Evidence from the Japan-US inventor survey. RIETI Discussion Paper Series 09-E -034.

Wieser, R. 2005. Research and development, productivity and spillovers: Empirical evidence at the firm level. *Journal of Economic Surveys* 19 (4): 587–621.

Zhen, Ni. 2018. Employment dynamics, firm performance and innovation persistence in the context of differentiated innovation types: Evidence from Luxembourg. PhD dissertation.

Zuniga, M.P., and D. Guellec. 2009. Who licenses out patents and why? Lessons from a business survey. STI Working Paper 2009/5, OECD.

The Choice of Comparable DMUs and Environmental Variables

John Ruggiero

1 Introduction

With economic production theory as a basis, Farrell (1957) showed how overall efficiency can be estimated relative to and decomposed into technical efficiency and allocative efficiency. Technical inefficiency is observed when a given production possibility is not on the isoquant. As a result, the unit is using too many inputs to produce the observed output, leading to excess costs and lower profits. Allocative inefficiency results when the firm uses the wrong mix of inputs given exogenous input prices. Farrell and Fieldhouse (1962) extended Farrell's earlier work by relaxing the assumption of constant returns to scale by allowing decreasing returns. In addition, linear programming was suggested as a methodology that could solve for inefficiency. Boles (1971) extended the models to variable returns to scale and provided computer programs to estimate the efficiency. Afriat (1972) provided the formulation for technical efficiency measurement that was consistent with data envelopment analysis (DEA) with variable returns to scale and the Free Disposal Hull model. Färe et al. (1994) provide a useful theoretical framework for the production economic approach to efficiency measurement.

DEA was introduced to the operations research literature by Charnes et al. (1978) to measure the technical efficiency of a given observed decision

J. Ruggiero (✉)
University of Dayton, Dayton, OH, USA
e-mail: jruggiero1@udayton.edu

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_5

making unit (DMU) assuming constant returns to scale for a multiple input, multiple output production correspondence. The model was extended by Banker et al. (1984) to allow variable returns to scale; solutions to the constant returns to scale and variable returns to scale models allowed a further decomposition into technical and scale efficiency components. The justification for focusing on technical efficiency instead of cost efficiency was the lack of input prices in public sector applications. In addition to estimates of technical efficiency, the DEA model provides apparent benchmarking capabilities. Given that frontier estimates are convex combinations of observed technically efficient DMUs, the resulting referent set provides comparisons for inefficient units. Secondary analysis could provide best practices that could possibly inform inefficient DMUs and provide a path to technical efficiency. Alternatively, one could interpret DEA as a nonparametric estimator of frontier production; in this case, any frontier point could serve as a relevant benchmark where the path to efficiency allowed increasing or decreasing discretionary inputs.

While the supposed lack of input and output prices in public sector applications provided a rationale to use DEA, applications to the public sector revealed another weakness. Unlike regression analysis, DEA did not have a way to include nondiscretionary variables. With respect to public sector applications, there is a vast literature showing that the production is a function not only of discretionary inputs, but also of nondiscretionary inputs. Here, we consider nondiscretionary inputs as factors that influence the amount of output but that are taken as given by DMU. Early contributions to the DEA literature included applications analyzing public sector education. Charnes et al. (1981) applied the constant returns to scale model to analyze program and managerial efficiency of Program Follow Through. In this paper, nondiscretionary factors (education of the mother, highest occupation of a family member, etc.) were included as discretionary inputs.

Bessent et al. (1982) analyzed the 167 elementary schools in the Houston Independent School District and used several nondiscretionary factors (e.g., percent of students paying full lunch price and percent of nonminority students) as discretionary inputs. Smith and Mayston (1987) illustrated DEA with English school authorities using a constant returns to scale model. While they correctly distinguished between discretionary and nondiscretionary inputs, both types of inputs were treated similarly in the DEA model. Färe et al. (1989) measured efficiency of Missouri schools and used discretionary inputs and standardized tests. An attempt was made to control for nondiscretionary factors by restricting the sample to a homogenous group of DMUs.

Thanassoulis and Dunstan (1994) analyze cohorts of students using DEA with targets for improvement. The DEA models used a pretest to control for prior attainment and a socioeconomic variable (percent of students not receiving free school lunches) as the discretionary inputs.

Beginning with the Coleman Report (1966), there has been strong evidence that socioeconomic variables are the most important factors determining educational outcomes. While quality teacher and other discretionary inputs can positively impact outcomes, the empirical evidence suggests that parental background and student characteristics have a bigger effect. Failure to control for the socioeconomic variables leads to biased estimates of frontier production and therefore, of technical efficiency.

Hanushek (1989) summarized approximately twenty years of educational production studies and concluded that differences in school spending do not explain variations in student performance. Family background, however, does explain the differences in outcomes. Hanushek further finds that students with wealthier and more educated parents perform better. Hanushek (1979, 1986) provides a useful foundation to analyze education as a production process whereby outcomes are function of school inputs and socioeconomic variables. In addition to Hanushek's work, Bridge et al. (1979) and Cohn and Geske (1990) provide useful discussions of the education production process.

Bradford et al. (1969) provided a two-stage model to analyze public sector production where intermediate outputs (e.g., instruction in mathematics, reading, etc.) are determined by school resources. In a second stage, the final outcomes of interest are functionally related to the intermediate outputs and the socioeconomic environment. Importantly for our work, these socioeconomic factors of production are nondiscretionary even in the long-run. For purposes of measuring efficiency, it is important therefore to properly control for the socioeconomic environment.

Theoretical extensions useful for analyzing educational production were made by Banker and Morey (1986) which allowed nondiscretionary inputs. Alternative models to control for the socioeconomic environment include Ray (1988, 1991) which used a second-stage ordinary least squares regression. McCarty and Yaisawarng (1993) extended this by using a Tobit regression in the second stage.[1] Ruggiero (1996, 1998) provided a conditional

[1]Simar and Wilson (2007) criticize the two-stage models and argue in favor of a bootstrapping approach. Banker and Natarajan (2008) and McDonald (2009) prove the consistency of the OLS estimator in the second stage. McDonald (2009) further shows that Tobit is not appropriate.

technology that does not assume convexity with respect to the nondiscretionary variables. Ruggiero (1996) extended the DEA model to allow benchmarking of a given DMU only to those who had an environment (defined by the level of the nondiscretionary input) no better than that DMU. With multiple nondiscretionary variables, however, the model does not allow comparison of DMUs that are better in some but not all factors. Ruggiero (1998) provided a three-stage model that develops an overall index of environmental harshness with a second-stage regression that weights the importance of each nondiscretionary input. In the third stage, the technology is conditional on the estimated second-stage environmental index.[2]

2 Data Envelopment Analysis

Assume that each of n DMUs uses a vector $X = (x_1, \ldots, x_m)$ of m discretionary inputs to produce a vector $Y = (y_1, \ldots, y_s)$ of s outputs. Input and output data for individual DMU j ($j = 1, \ldots, n$) are represented by $X_j = (x_{1j}, \ldots, x_{mj})$ and $Y_j = (y_{1j}, \ldots, y_{sj})$, respectively. Assuming variable returns to scale, the empirical production possibility set is defined as:

$$\begin{aligned} \tau_v = \{(Y,X) : & \sum_{j=1}^{n} \lambda_j y_{kj} \geq y_k,\ k = 1, \ldots, s; \\ & \sum_{j=1}^{n} \lambda_j x_{lj} \leq x_l,\ l = 1, \ldots, m; \\ & \sum_{j=1}^{n} \lambda_j = 1; \\ & \lambda_j \geq 0,\ j = 1, \ldots, n\}. \end{aligned} \tag{1}$$

The technology τ_v is characterized by variable returns to scale given the convexity constraint. It is assumed that any convex combination of observed production (and hence all observed production) is feasible.

[2]Ruggiero (2000) extended the conditional model to estimate returns to scale. Brennan et al. (2013) and Johnson and Ruggiero (2014) extended the model to measure productivity. Extensions to measuring school costs and allocative efficiency were provided by Ruggiero (2001) and Haelermans and Ruggiero (2013), respectively.

Measurement of technical efficiency in DEA uses observed production possibilities to construct the observed production possibility set; estimates of efficiency are (usually) obtained using the Farrell measure with either an input or output orientation. We first define the output-oriented DEA measure of technical efficiency as the reciprocal of the distance function:

Definition: $\mathrm{TE}^{O}(Y_i, X_i) = (\max\{\theta : (\theta Y_i, X_i) \in \tau_v)\})^{-1} \leq 1$ is the output-oriented measure of technical efficiency for DMU i $(i = 1, \ldots, n)$.

The nonparametric estimate of the technical efficiency for DMU i $(i = 1, \ldots, n)$ is obtained in the solution of the following linear programming model:

$$
\begin{aligned}
&\left[\mathrm{TE}^{O}(Y_i, X_i)\right]^{-1} = \mathrm{Max}\ \theta \\
&\text{s.t.} \\
&\sum_{j=1}^{n} \lambda_j y_{kj} \geq \theta y_{ki},\ k = 1, \ldots, s; \\
&\sum_{j=1}^{n} \lambda_j x_{lj} \leq x_{li},\ l = 1, \ldots, m; \\
&\sum_{j=1}^{n} \lambda_j = 1; \\
&\lambda_j \geq 0,\ j = 1, \ldots, n.
\end{aligned}
\tag{2}
$$

For a given DMU, the output-oriented model identifies the maximum equiproportional expansion of observed output possible holding input levels no greater than the observed levels. The solution to (2) for DMU i provides not only the measure of technical efficiency but also a convex combination with weights $(\lambda_{i1}^*, \ldots, \lambda_{iN}^*) = (\lambda_1^*, \ldots, \lambda_N^*)$ that can serve as a benchmark.

Alternatively, one could use an input-oriented model that seeks the maximum equiproportional reduction in observed inputs consistent with the observed output levels. We define the DEA input-oriented measure of efficiency as:

Definition: $\mathrm{TE}^{I}(Y_i, X_i) = (\min\{\theta : (Y_i, \theta X_i) \in \tau_v)\}) \leq 1$ is the input-oriented measure of technical efficiency for DMU i $(i = 1, \ldots, n)$.

The DEA estimate of technical efficiency for DMU i $(i = 1, \ldots, n)$ can be obtained in the solution of the following linear programming model:

$$
\begin{aligned}
&\mathrm{TE}^I(Y_i, X_i) = \mathrm{Min}\ \theta \\
&\text{s.t.} \\
&\sum_{j=1}^{n} \lambda_j y_{kj} \geq y_{ki},\ k = 1, \ldots, s; \\
&\sum_{j=1}^{n} \lambda_j x_{lj} \leq \theta x_{li},\ l = 1, \ldots, m; \\
&\sum_{j=1}^{n} \lambda_j = 1; \\
&\lambda_j \geq 0,\ j = 1, \ldots, n.
\end{aligned} \tag{3}
$$

Similar to the output-oriented model, (3) assumes that all observed production and any convex combination of observed production are feasible. And, the solution of the model provides not only an estimate of technical efficiency but also a relevant benchmark with $(\lambda_{i1}^*, \ldots, \lambda_{iN}^*) = (\lambda_1^*, \ldots, \lambda_N^*)$ for each DMU $i\,(i = 1, \ldots, n)$.

For illustrative purposes, we provide a simple example where four DMUs are observed producing one output (y) using one input (x) We further assume that frontier production is given by $y = x^{0.5}$. Data for the four observations are shown below:

DMU	Input (x)	Output (y)
A	1	1
B	4	2
C	9	3
D	6	1.5

The data are shown in Fig. 1.

Based on the known production frontier, we observe that *A*, *B* and *C* are technically efficient in both the input- and output-oriented models. DMUs *A*, *B* and *C* are each observed producing the maximum output given the observed input and simultaneously employing the least input given observed output. DMU *D*, however, is technically inefficient. *D* is observed producing 1.5 units of output; according to the known production frontier, the minimum level of input necessary to produce this output is 2.25. In the solution of (3), we estimate $\mathrm{TE}^I(Y_D, X_D) = 0.417$ with a benchmark consisting of an equally weighted combination of *A* and *B*. We note that the referent frontier convex combination (2.5, 1.5) is not truly efficient because a piecewise linear is used to approximate the true frontier.

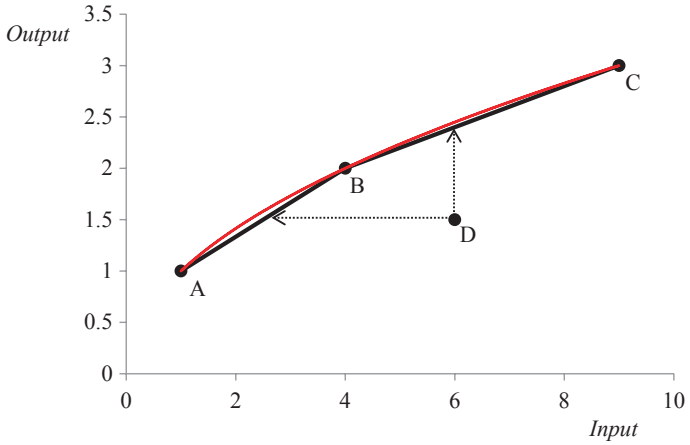

Fig. 1 DEA and benchmarking

For the output-oriented model, we know that D should be able to produce an output of $\sqrt{6} \approx 2.45$. D is observed producing an output of only 1.5, leading to a true output-oriented efficiency of approximately 0.61. The solution of (2) for DMU D leads to an estimate of technical efficiency of $\mathrm{TE}^{O}(Y_D, X_D) = 0.625$ using a referent convex combination with weights on B and C of 0.6 and 0.4, respectively.

DEA purportedly provides benchmarking for inefficient firms. In the case where there are efficient frontier observations using the same input levels (in the case of the output-oriented model), we do obtain benchmarks that are relatively efficient. In the case illustrated in Fig. 1, however, the benchmark is a convex combination of relatively efficient DMUs. Presumably, DMUs B and C, for example, are doing something different and better than DMU D. If, for example, B, C and D all employ different managerial styles, we really only know that the managerial styles of B and C are better than the managerial style of D. But it is not clear what a convex combination of the managerial styles of B and C even means. Alternatively, one could argue that A, B and C are all useful benchmarks for D conditional on which input and output combination D would prefer assuming it became technically efficient.

The DEA models envelop the observed data with piecewise linear facets to provide an estimate of frontier production in deterministic models where all deviations are due to inefficiency. In these DEA models, it is assumed that all inputs and outputs are discretionary. If production is characterized by nondiscretionary inputs, then the equiproportional reduction in these factors of production is inconsistent with the assumed model. In other words, the benchmark for DMUs with nondiscretionary inputs should be to

feasible production possibilities holding the nondiscretionary inputs at the observed level. In the next section, we extend the analysis to the nondiscretionary inputs.

3 DEA in the Short-Run

In standard economics, the short-run is defined as the period of time where at least one factor of production is considered fixed. In the long-run, all inputs are variable. Hence, in the short-run, it may not be possible to reduce the amount of some input, say, capital. Based on the definitions above, this will not have any effect on the output-oriented estimates of technical efficiency because technical efficiency is defined holding all of the inputs fixed. Furthermore, the input-oriented model might still be appropriate if the wasted amount of the fixed input can be freely disposed. Assuming the technology is characterized by convexity in the long-run, then the input-oriented models will provide a feasible frontier possibility. Some capital like machines and equipment could be held idle. However, in some cases, the amount of capital cannot be varied (number of plants, for example) in the short-run and hence, can be considered nondiscretionary in the short-run and discretionary in the long-run.

We now consider production of the technology in the short-run with fixed inputs. Using the notation above, we now assume that each of n DMUs uses a vector $X = (x_1, \ldots, x_m)$ of m discretionary inputs given a vector $Z = (z_1, \ldots, z_p)$ of p fixed inputs to produce a vector $Y = (y_1, \ldots, y_s)$ of s outputs. The fixed input data for DMU j ($j = 1, \ldots, n$) are given by $Z_j = (z_{1j}, \ldots, z_{pj})$. Given the assumption of convexity, the empirical production possibility set is now defined as:

$$\begin{aligned}\tau_v^1 = \{(Y, X, Z) : & \sum_{j=1}^{n} \lambda_j y_{kj} \geq y_k,\ k = 1, \ldots, s; \\ & \sum_{j=1}^{n} \lambda_j x_{lj} \leq x_l,\ l = 1, \ldots, m; \\ & \sum_{j=1}^{n} \lambda_j z_{qj} \leq z_q,\ q = 1, \ldots, p; \\ & \sum_{j=1}^{n} \lambda_j = 1; \\ & \lambda_j \geq 0,\ j = 1, \ldots, n\}.\end{aligned} \tag{4}$$

Here, the technology is equivalent to (1) where we make a distinction between the fixed and variable inputs. Using this technology, we now redefine the definition of the output-oriented measure of efficiency relative to the technology in (4):

Definition: $\mathrm{TE}^O(Y_i, X_i, Z_i) = (\max\{\theta : (\theta Y_i, X_i, Z_i) \in \tau_v^1)\})^{-1} \leq 1$ is the output-oriented measure of technical efficiency for DMU $i\,(i = 1, \ldots, n)$.

Given the symmetric treatment of the variable and fixed inputs in the output-oriented model, namely that all inputs are held fixed in the estimation of efficiency, the estimate of technical efficiency for DMU i $(i = 1, \ldots, n)$ is obtained in the solution of the following linear program:

$$
\begin{aligned}
&\left[\mathrm{TE}^O(Y_i, X_i, Z_i)\right]^{-1} = \mathrm{Max}\ \theta \\
&\text{s.t.} \\
&\sum_{j=1}^{n} \lambda_j y_{kj} \geq \theta y_{ki},\ k = 1, \ldots, s; \\
&\sum_{j=1}^{n} \lambda_j x_{lj} \leq x_{li},\ l = 1, \ldots, m; \\
&\sum_{j=1}^{n} \lambda_j z_{qj} \leq z_{qi},\ q = 1, \ldots, p; \\
&\sum_{j=1}^{n} \lambda_j = 1; \\
&\lambda_j \geq 0,\ j = 1, \ldots, n.
\end{aligned}
\tag{5}
$$

Essentially, to estimate output-oriented efficiency, all inputs are treated fixed and the solution of (5) for each DMU provides a measure of efficiency as the reciprocal of the maximum equiproportional expansion of all outputs. Further, the model provides benchmarking information with $(\lambda_{i1}^*, \ldots, \lambda_{iN}^*) = (\lambda_1^*, \ldots, \lambda_N^*)$ for each DMU $i\,(i = 1, \ldots, n)$.

We now define the DEA input-oriented measure of technical efficiency relative to the technology (4):

Definition: $\mathrm{TE}^I(Y_i, X_i, Z_i) = (\min\{\theta : (Y_i, \theta X_i, Z_i) \in \tau_v^1)\}) \leq 1$ is the input-oriented measure of technical efficiency for DMU $i\,(i = 1, \ldots, n)$.

This definition of technical efficiency in the case of fixed inputs modifies the original definition by considering the equiproportional reduction in variable inputs holding outputs and fixed inputs at the observed level. Banker and Morey (1986) provided the estimator of technical efficiency for DMU $i\,(i = 1, \ldots, n)$ as the solution of the following linear programming model:

$$
\begin{aligned}
&\mathrm{TE}^I(Y_i, X_i, Z_i) = \mathrm{Min}\ \theta \\
&\text{s.t.} \\
&\sum_{j=1}^{n} \lambda_j y_{kj} \geq y_{ki},\ k = 1, \ldots, s; \\
&\sum_{j=1}^{n} \lambda_j x_{lj} \leq \theta x_{li},\ l = 1, \ldots, m; \\
&\sum_{j=1}^{n} \lambda_j z_{qj} \leq z_{qi},\ q = 1, \ldots, p; \\
&\sum_{j=1}^{n} \lambda_j = 1; \\
&\lambda_j \geq 0,\ j = 1, \ldots, n.
\end{aligned}
\tag{6}
$$

The estimator for linear programming model (6) allows convex combinations of production possibilities defined on the outputs, variable inputs and fixed inputs. Essentially, the Banker and Morey (1986) model projects inefficient points to the production frontier using an equiproportional reduction in observed variable inputs, holding outputs and fixed inputs at the observed values of the DMU under analysis.

The Banker and Morey (1986) model is illustrated in Fig. 2, where it is assumed that four DMUs *A–D* are observed producing the same level of output $y_1 = 10$ using one discretionary input x_1 and one fixed input z_1. Efficient production for this example is given by $y_1 = x_1^{0.3} z_1^{0.5}$. Observed production data are shown in the following table:

DMU	x_1	z_1
A	4.61	40
B	11.53	20.06
C	20	16.57
D	20	40

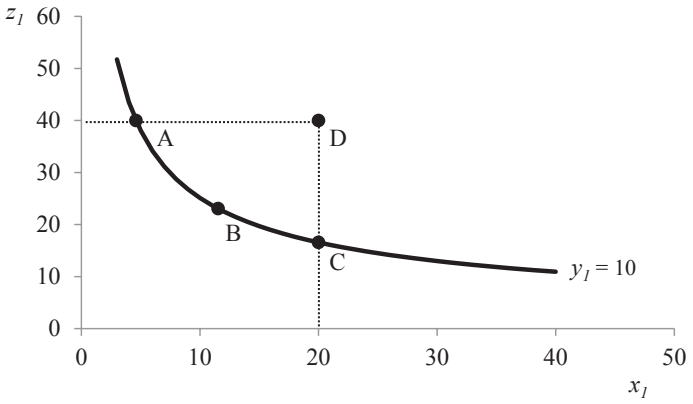

Fig. 2 DEA with a nondiscretionary input

The data are illustrated in Fig. 2 with the discretionary input x_1 on the horizontal axis and the nondiscretionary input z_1 on the vertical axis. Based on the production function we observe *A*–*C* are all technically efficient producing the observed output with the fewest inputs. In the traditional sense where z_1 represents capital that is fixed in the short-run, the technical efficiency measure $\mathrm{TE}^I(Y_D, X_D) = 0.577$. Viewing technical efficiency in the long-run sense, DMU *D* would be able to reduce both x_1 and z_1 to approximately 57.7% of its current input use and still produce $y_1 = 10$. The benchmark in this case would be DMU *B*.

As pointed out by Banker and Morey (1986), this expansion path is not possible in the short-run. Instead, one could estimate (6) and seek the maximum reduction in input x_1 while holding z_1 fixed at $z_{1D} = 40$. The relevant benchmark in this case would be DMU *A* which is observed having the same amount of the fixed input $z_{1A} = 40$ and using less of the discretionary input x_1 to produce the same output level. As a result, $\mathrm{TE}^I(Y_D, X_D, Z_D) = 4.06/20 = 0.23$, which implies DMU *D* could reduce its discretionary input to 23% of its current level and still produce $y_1 = 10$ holding $z_{1D} = 40$.

4 DEA with Nondiscretionary Variables

Many production processes involve factors of production that are nondiscretionary even in the long-run. For example, in public service provision, socioeconomic variables beyond the control of the DMU influence the amount of resources necessary to produce a given amount of output. It is well known in the education literature that socioeconomic status plays a large role in the

success of the students. As a result, it is necessary to control for the socioeconomic environment to measure a given DMUs efficiency and relevant benchmarks. The environmental variables are nondiscretionary variables and will be treated as such in the development of the model.

We now consider the production technology in the presence of a nondiscretionary variable. We now assume that each of n DMUs uses a vector $X = (x_1, \ldots, x_m)$ of m discretionary inputs given an environmental variable z_1 to produce a vector $Y = (y_1, \ldots, y_s)$ of s outputs. The environmental variable for DMU j $(j = 1, \ldots, n)$ is given by z_{1j}. We assume that a more favorable environment is represented by a higher value of the environmental variable z_1. The empirical production possibility set conditional on the environmental variable z defined by Ruggiero (1996)[3] is:

$$\begin{aligned} \tau_v^1(z) = \{(Y, X, z) : & \sum_{j=1}^{n} \lambda_j y_{kj} \geq y_k, \; k = 1, \ldots, s; \\ & \sum_{j=1}^{n} \lambda_j x_{lj} \leq x_l, \; l = 1, \ldots, m; \\ & \sum_{j=1}^{n} \lambda_j = 1; \\ & \lambda_j (z - z_j) \geq 0, \; j = 1, \ldots, n; \\ & \lambda_j \geq 0, \; j = 1, \ldots, n\}. \end{aligned} \tag{7}$$

In this representation, a constraint $\lambda_j(z - z_j) \geq 0$ is added for each DMU j $(j = 1, \ldots, n)$ to insure that any DMU with a more favorable environment (i.e., $z_j > z$) is disallowed from serving as a referent benchmark. Hence, the production possibility set is conditional on the environmental variable. Using this technology, we now redefine the definition of the output-oriented measure of efficiency relative to the technology in (7):

Definition: $\mathrm{TE}^O(Y_i, X_i, z_i) = (\max\{\theta : (\theta Y_i, X_i, z_i) \in \tau_v^1(z_i)\})^{-1} \leq 1$ is the output-oriented measure of technical efficiency for DMU i $(i = 1, \ldots, n)$.

[3]We use the notation of Podinovski (2005) to define the production possibility set.

Unlike the Banker and Morey (1986) model, discretionary inputs and discretionary outputs are treated similarly to the standard DEA model with the exception that projection to the output isoquant is conditional on the environmental variable. DMUs that have a more favorable environment are excluded as benchmarks in the evaluation of the DMU under analysis. An estimate of technical efficiency in the output-oriented model for DMU i $(i = 1, \ldots, n)$ is obtained in the solution of the following linear program developed by Ruggiero (1996):

$$
\begin{aligned}
&\left[\mathrm{TE}^{O}(Y_i, X_i, z_i)\right]^{-1} = \mathrm{Max}\ \theta \\
&\text{s.t.} \\
&\sum_{j=1}^{n} \lambda_j y_{kj} \geq \theta y_{ki},\ k = 1, \ldots, s; \\
&\sum_{j=1}^{n} \lambda_j x_{lj} \leq x_{li},\ l = 1, \ldots, m; \\
&\sum_{j=1}^{n} \lambda_j = 1; \\
&\lambda_j (z_i - z_j) \geq 0, j = 1, \ldots, n; \\
&\lambda_j \geq 0,\ j = 1, \ldots, n.
\end{aligned}
\tag{8}
$$

Efficiency is estimated as the inverse of the equiproportional expansion of all discretionary outputs holding discretionary inputs fixed conditional on the environmental variable. In the evaluation of a given DMU i, the optimal weight $\lambda_{ij}^{*} = 0$ for any DMU j that has a more favorable environment (i.e., $z_j \geq z_i$).

We can also define an input-oriented measure of technical efficiency relative to the variable returns to scale technology defined in (7):

Definition: $\mathrm{TE}^{I}(Y_i, X_i, z_i) = \min\{\theta : (Y_i, \theta X_i, z_i) \in \tau_v^1(z_i)\} \leq 1$ is the input-oriented measure of technical efficiency for DMU i $(i = 1, \ldots, n)$.

Here, benchmarks are identified from the equiproportional reduction in discretionary inputs for a given level of outputs conditional on the environment. Estimation of technical efficiency conditional on the environment was provided in Ruggiero (1996) for DMU i $(i = 1, \ldots, n)$ as the solution to the following linear programming model:

$$
\begin{aligned}
&\mathrm{TE}^I(Y_i, X_i, z_i) = \text{Min } \theta \\
&\text{s.t.} \\
&\sum_{j=1}^{n} \lambda_j y_{kj} \geq y_{ki},\ k = 1, \ldots, s; \\
&\sum_{j=1}^{n} \lambda_j x_{lj} \leq \theta x_{li},\ l = 1, \ldots, m; \\
&\sum_{j=1}^{n} \lambda_j = 1; \\
&\lambda_j (z_i - z_j) \geq 0, j = 1, \ldots, n; \\
&\lambda_j \geq 0,\ j = 1, \ldots, n.
\end{aligned}
\tag{9}
$$

The models (8) and (9) were shown to work well using simulated data in Ruggiero (1998) when one variable captured the effect of the environment. In cases where there are multiple environmental variables, the model is unable to individually weight the importance of each environmental variable without a relatively large number of observations. Ruggiero (1998) proposed a three-stage model where only discretionary variables are used in the first stage. The second-stage model used regression analysis to derive an overall environmental index that was incorporated into a third-stage model using either (8) or (9). See Estelle et al. (2010) for a further discussion.

We illustrate the conditional technology with a simulation where we assume one output y_1 is produced using one discretionary input x_1 and one environmental variable z_1 where efficient production is given by $y_1 = z_1 x_1^{0.4}$. We assume three different values for z_1: 1, 1.5 and 2. For each level of z_1 we vary x_1 from 0.5 to 11.5 in increments of 0.5. For these 69 observations, we calculate observed production at the efficient level according to the production function. Additionally, we generate three additional points presented in the following table:

DMU	x_1	z_1	y_1
A	6	1.5	2.5
B	3.5861	1.5	2.5
C	6	1.5	3.0715

The data are illustrated in Fig. 3.

DMU *A* is the only inefficient DMU. DMUs *B* and *C* are the appropriate benchmarks for DMU *A*; we note that both DMUs *B* and *C* are observed

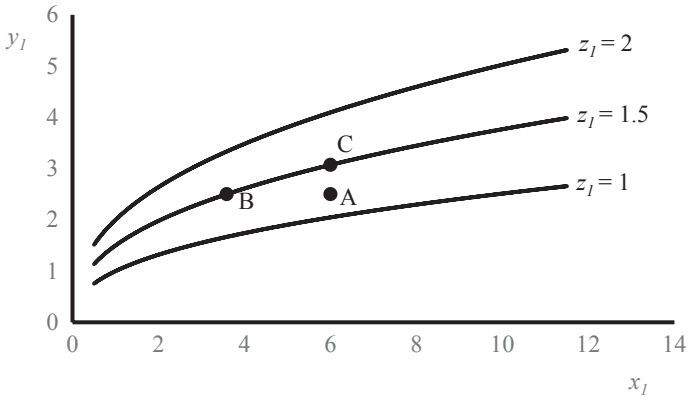

Fig. 3 Frontiers conditional on z_1

producing on the production frontier associated with the moderate environment level with $z_1 = 1.5$, i.e., the level of the environmental variable of DMU *A*. DMU *B* is the input-oriented benchmark for DMU *A*; this reveals that the technical efficiency of *A* is only 59.77% efficient: DMU *A* could reduce its discretionary input x_1 to 59.77% of its observed input level and still produce the observed output $y_1 = 2.5$. Likewise, DMU *C* is the output-oriented benchmark for DMU *A*; given $x_1 = 6$, DMU *A* could increase observed output from 2.5 to 3.0715 leading to an output-oriented level of technical efficiency of 81.39%.

We now consider the solution to the Banker and Morey (1986) input-oriented model (6). In the solution of (6) for DMU *A*, the referent set consists of three observed efficient points $(x_1, z_1, y_1) = (1.5, 1, 1.1760)$, (5, 2, 3.8073) and (5.5, 2, 3.9553) with weights of 0.5, 0.444 and 0.056, respectively. The resulting convex combination is $(x_1, z_1, y_1) = (3.278, 1.5, 2.5)$. But this convex combination is not feasible given that DMU *B* is the efficient referent point with $x_1 = 3.5861$. As a result, the estimated efficiency of DMU *A* (0.54635) is biased downward from the true efficiency measured to be 0.5977. This results because estimator (6) wrongly imposes convexity with respect to the nondiscretionary input z_1. Estimating the model with (9) produces the correct results; DMU *A* is only 0.5977 efficient with a benchmark consisting of DMU *B* with a weight of 1.

We also consider the output-oriented solution using the simulated data. Using estimator (5), benchmark for DMU *A* consists of an equally weighted convex combination of $(x_1, z_1, y_1) = (3, 1, 1.5519)$ and (9, 2, 4.8165) which implies that DMU *A* should be able to produce $y_1 = 3.1842$ given $x_1 = 6$ and $z_1 = 1.5$. However, this is not feasible. Given the production function used

in the data-generating process, we know that the maximum possible output that can be obtained from $x_1 = 6$ and $z_1 = 1.5$ is $y_1 = 3.0715$ (i.e., DMU *C*). The convexity assumption leads to an infeasible projection and hence, an incorrect benchmark. Estimating the efficiency using (9) produces a feasible projection, the correct measure of technical efficiency and the appropriate benchmark of DMU *C*. Of course, the precise benchmark was generated because the true benchmarks *B* and *C* were included in the data set.

5 Analysis of Educational Production

To illustrate the effect of using the alternative models, we consider an application of DEA to analyze performance of Ohio school districts. Data for this application were used in Johnson and Ruggiero (2014). In this application, we consider only data for the 2007–2008 school year. In Ohio, thirty measures of performance (primarily results on standardized tests) are collected and aggregated into one measure of overall outcome measure. We note that the measure is highly correlated with each of the 30 individual performance measures. We use this measure for the output in the DEA models.

Four expenditure per pupil measures were chosen as the discretionary inputs: administrative, instructional, building operation and pupil support. Given that expenditures are dependent on prices, all expenditures were deflated by an index of first-year teacher salaries (using data from school year 2004–2005.) We also consider the percent of students not in poverty as a nondiscretionary input. Descriptive statistics for the 204 school districts are reported in Table 1.

By far, instructional expenditures per pupil represent the largest share of total expenditures per pupil. On average, nearly 56.9% of all expenditures per pupil are spent on instruction (primarily on teachers). On average, school districts spend approximately 12.7% of total operational expenditures on administration.

For comparative purposes, we estimate the fixed input model of Banker and Morey (6) and the conditional convexity model of Ruggiero (9) to the Ohio school district data. The results of the estimation are reported in Table 2.

Both of the models provided similar estimates. We know that the Banker and Morey model will always produce technical efficiency estimates no greater than the Ruggiero model. On average, the Banker and Morey estimated technical efficiency to be 0.81 with a standard deviation of 0.105. The Ruggiero model estimated average efficiency to be 0.848, a difference of 0.039. The correlation (rank correlation) between the two estimators

Table 1 Descriptive statistics Ohio school districts ($N = 604$)

Variable	Mean	Std. dev.	Min.	Max.
Output				
State Performance Index	95.824	6.313	70.000	109.700
Discretionary inputs				
Instructional expenditures	$5018	$723	$3396	$9721
Administrative expenditures	$1123	$316	$576	$4295
Building operations expenditures	$1786	$442	$943	$6164
Pupil support expenditures	$897	$257	$358	$3192
Nondiscretionary input				
Percent of students not in poverty	96.96	3.35	76.97	100.00

Expenditures are all measured per pupil. Calculations by Author

Table 2 Summary of results

Estimator	Mean	Std. dev.	Min.	Max.
Banker and Morey	0.809	0.105	0.365	1.000
Ruggiero	0.848	0.110	0.414	1.000
Difference	0.039	0.049	0.000	0.298
Correlation between estimators				
Correlation	0.895			
Rank correlation	0.887			

All estimates by Authors. Correlations reported are the Pearson and Spearman correlation coefficients between the input-oriented Banker and Morey estimator (6) and the input-oriented Ruggiero estimator (9)

was 0.895 (0.887), implying that the assumption of convexity with respect to the nondiscretionary input in the Banker and Morey model is not that problematic. The largest difference between the two estimators was 0.298, suggesting the possibility that the Banker and Morey model overestimate technical inefficiency in some cases.

We turn now to benchmarking in the two models. We'll focus on a specific case where the two models produced widely different results. In particular, we choose DMU 382 which was estimated to be 66.8% efficient using the Banker and Morey estimator and 87.8% efficient using the Ruggiero estimator, leading to a difference of 0.210 in the estimated efficiency. In Table 3, we report the benchmark obtained from the solution of (6) which assumes convexity with respect to all inputs. In Table 4, we report the benchmark from the solution of the conditional convexity model (9).

Included in both tables are the variables used in the DEA models. Consider first the Banker and Morey results presented in Table 3. Five DMUs comprise the reference set used as a benchmark for DMU 382.

Table 3 Benchmark for DMU 382 using Banker and Morey model

Variable	DMU 382	Benchmark DMUs				
		DMU 32	DMU 44	DMU 136	DMU 264	DMU 500
Output						
State Performance Index	94.6	72.1	76.0	97.1	100.7	92.7
Discretionary inputs						
Instructional expenditures	$6351	$5794	$6040	$5789	$3553	$3909
Administrative expenditures	$1248	$1026	$1892	$1177	$720	$599
Building operations expenditures	$3046	$1705	$1791	$1741	$1347	$2165
Pupil support expenditures	$831	$788	$719	$867	$430	$581
Nondiscretionary input						
Students not in poverty (%)	92.56	82.64	80.21	81.81	97.24	93.82
Weight		0.157	0.020	0.113	0.622	0.087

All calculations by Author. Expenditures are measured per pupil

Table 4 Benchmark for DMU 382 using Ruggiero model

Variable	DMU 382	Benchmark DMUs		
		DMU 250	DMU 498	DMU 505
Output				
State Performance Index	94.6	93.5	89.6	97.7
Discretionary inputs				
Instructional expenditures	$6351	$4639	$5782	$5259
Administrative expenditures	$1248	$1353	$936	$1071
Building operations expenditures	$3046	$1900	$1807	$2584
Pupil support expenditures	$831	$472	$473	$973
Nondiscretionary input				
Students not in poverty (%)	92.56	92.54	92.33	90.361
Weight		0.216	0.271	0.513

All calculations by Author. Expenditures are measured per pupil

DMU 264 has the highest weight (0.622) used to evaluate DMU 382. We note that DMU 264 has 97.24% of their students not in poverty. This is much larger than DMU 382, which has more poverty with only 92.56% of students not in poverty. Compared to DMU 382, this referent DMU had a much higher State Performance Index (100.7 vs. 94.6) and much lower per pupil expenditures on instruction ($3553 vs. $6351), administration ($720 vs. $1248), building operations ($1347 vs. $3046) and pupil support ($430 vs. $831). Given that DMU 264 has a much more favorable environment, it is not clear if the much lower input levels and higher outcome level should be attributed to the more favorable environment or to more efficient use of resources. Another DMU in the reference set, DMU 500 with a weight of only 0.087, also has much lower input levels and a more favorable environment with a lower percent of students in poverty. The convex combination also includes three DMUs (32, 44 and 136) with a total weight of 0.290 that have a much worse environment. Of these, DMU 136 was able to achieve much better, a better outcome (97.1 vs. 94.6) than DMU 382 with lower expenditures per pupil in instruction, administration and building operations but higher expenditures in pupil support. Overall, it is not clear whether the estimated inefficiency is caused by better-performing school districts in the referent set or is biased downward by inclusion of districts with a much more favorable environment.

Table 4 reports the benchmark using estimator (9) introduced by Ruggiero (1996). The referent benchmark for DMU 382 consists of a convex combination of three different DMUs (250, 498 and 505) than the ones selected by the Banker and Morey model. Here, all three DMUs have a higher percentage of students in poverty than DMU 382. As a result, unlike

in the Banker and Morey model, the results cannot be attributed to including districts with a more favorable environment in the referent set. In terms of the output measure, DMU 505 (with a weight of 0.513) had a higher performance index than DMU 382 (97.7 vs. 94.6) while the other DMUs had a lower performance index value. However, with two exceptions (DMU 250's administrative expenditures per pupil and DMU 505's pupil support expenditures per pupil), all per pupil expenditures were lower than DMU 382's. In terms of benchmarking, it is clear that differences between DMU 382 and the referent convex combination arise from better performance and not a better environment.

6 Conclusions

In this chapter, we presented alternative DEA models that have been used in applications where environmental variables are inputs into the production process. Many applications have used the standard DEA model and have treated the uncontrollable inputs as discretionary in an input-oriented framework. If in fact these factors cannot be reduced, then the resulting measures of efficiency are biased and the resulting benchmarks are not appropriate. Banker and Morey (1986) provided the first model to incorporate fixed factors of production. Their model is appropriate as an application to short-run production where the DEA technology axioms are appropriate for long-run production. In this case, the Banker and Morey (1986) model properly estimates efficiency relative to the possible reduction in discretionary inputs and produces useful benchmarking information by holding discretionary inputs fixed.

In public sector production, however, the socioeconomic variables are fixed even in the long-run. There is no theoretical reason why convexity should hold for these factors of production, and hence, the Banker and Morey (1986) model can produce infeasible projections with the maintained assumptions. Alternatively, Ruggiero (1996) introduced a DEA model that controls for nondiscretionary variables by restricting possible benchmarks to include only those that have an environment no better than the unit under analysis. The model assumes that there is no correlation between efficiency and the nondiscretionary socioeconomic environment which would lead to an endogeneity problem. In this case, one could incorporate a multiple stage model incorporating two-stage least squares for example but this would require obtaining valid instruments.

References

Afriat, S.N. 1972. Efficiency estimation of production functions. *International Economic Review* 13: 568–598.

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

Banker, R., and R. Morey. 1986. Efficiency analysis for exogenous fixed inputs and outputs. *Operations Research* 34: 513–521.

Banker, R., and R. Natarajan. 2008. Evaluating contextual variables affecting productivity using data envelopment analysis. *Operations Research* 56: 48–58.

Bessent, A., E.W. Bessent, J. Kennington, and B. Reagan. 1982. An application of mathematical programming to assess productivity in the Houston independent school district. *Management Science* 28: 1335–1366.

Boles, J. 1971. *The 1130 Farrell efficiency system—Multiple products, multiple factors*. Berkeley: Giannini Foundation of Agricultural Economics.

Bradford, D., R. Malt, and W. Oates. 1969. The rising cost of local public services: Some evidence and reflections. *National Tax Journal* 22: 185–202.

Brennan, S., C. Haelermans, and J. Ruggiero. 2013. Non parametric estimation of education productivity incorporating nondiscretionary inputs with an application to Dutch schools. *European Journal of Operational Research* 234: 809–818.

Bridge, R., Judd, C., and P. Moock, P. 1979. *The determinants of educational outcomes*. Cambridge, MA: Ballinger Publishing.

Charnes, A., W.W. Cooper, and E. Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2: 429–444.

Charnes, A., W.W. Cooper, and E. Rhodes. 1981. Evaluating program and managerial efficiency: An application of data envelopment analysis to program follow through. *Management Science* 27: 668–697.

Cohn, E., and T. Geske. 1990. *The economics of education*. Oxford, UK: Pergamon Press.

Coleman, J., E. Campbell, C. Hobson, F. McPartland, A. Mood, F. Weinfeld, and R. York. 1966. *Equality of educational opportunity*. Washington, DC: Government Printing Office.

Estelle, S., A. Johnson, and J. Ruggiero. 2010. Three-stage DEA models for incorporating exogenous inputs. *Computers & Operations Research* 37: 1087–1090.

Färe, R., S. Grosskopf, and C.A.K. Lovell. 1994. *Production frontiers*. New York, NY: Cambridge University Press.

Färe, R., S. Grosskopf, and W. Weber. 1989. Measuring school district performance. *Public Finance Quarterly* 17: 409–428.

Farrell, M.J. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society* 120: 253–281.

Farrell, M.J., and M. Fieldhouse. 1962. Estimating efficient production functions under increasing returns to scale. *Journal of the Royal Statistical Society* 125: 252–267.

Haelermans, C., and J. Ruggiero. 2013. Estimating technical and allocative efficiency in the public sector: A nonparametric analysis of Dutch schools. *European Journal of Operational Research* 227: 174–181.

Hanushek, E. 1979. Conceptual and empirical issues in the estimation of educational production functions. *The Journal of Human Resources* 14: 351–388.

Hanushek, E. 1986. The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature* 24: 1141–1177.

Hanushek, E. 1989. The impact of differential expenditures on school performance. *Educational Researcher* 18: 45–62.

Johnson, A., and J. Ruggiero. 2014. Nonparametric measurement of productivity and efficiency in education. *Annals of Operations Research* 221: 197–210.

McCarty, T., and S. Yaisawarng. 1993. Technical efficiency in New Jersey school districts. In *The measurement of productive efficiency*, ed. H.O. Fried, C.A.K. Lovell, and S.S. Schmidt, 271–287. New York: Oxford University Press.

McDonald, J. 2009. Using least squares and tobit in second stage DEA efficiency analyses. *European Journal of Operational Research* 197: 792–798.

Podinovski, V. 2005. Selective convexity in DEA models. *European Journal of Operational Research* 161: 552–563.

Ray, S. 1988. Data envelopment analysis, nondiscretionary inputs and efficiency: An alternative interpretation. *Socio-Economic Planning Sciences* 22: 167–176.

Ray, S. 1991. Resource-use efficiency in public schools: A study of Connecticut data. *Management Science* 37: 1620–1628.

Ruggiero, John. 1996. On the measurement of technical efficiency in the public sector. *European Journal of Operational Research* 90: 553–565.

Ruggiero, John. 1998. Non-discretionary inputs in data envelopment analysis. *European Journal of Operational Research* 111: 461–469.

Ruggiero, John. 2000. Nonparametric estimation of returns to scale in the public sector with an application to the provision of educational services. *Journal of the Operational Research Society* 51: 906–912.

Ruggiero, John. 2001. Determining the base cost of education: An analysis of Ohio school districts. *Contemporary Economic Policy* 19 (3): 268–279.

Simar, L., and P. Wilson. 2007. Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics* 136: 31–64.

Smith, P., and D. Mayston. 1987. Measuring efficiency in the public sector. *Omega* 15: 181–189.

Thanassoulis, E., and P. Dunstan. 1994. Guiding schools to improved performance using data envelopment analysis: An illustration with data from a local education authority. *Journal of the Operational Research Society* 45: 1247–1262.

Data Envelopment Analysis with Alternative Returns to Scale

Subhash C. Ray

1 Introduction

Performance evaluation in any decision-making situation involves comparing the outcome from the decision actually made with what is deemed to be the most preferred outcome within the constraints of the decision-making problem. To evaluate the performance of a firm producing a single output from a specific bundle of inputs, one compares its actual output with the maximum producible quantity from the bundle of inputs it is using. In textbook economics, the production function defines the maximum output producible from any given bundle of inputs. The actual output may fall below this maximum due to inefficiency. In practice, there is no readily available scientific formula showing the maximum output from a given input and the production function has to be constructed from observed input-output data.

The common practice in empirical research is to start with an explicit functional specification of the production function and to use regression to estimate the parameters of the model. However, the two-sided residuals imply that that for some observations, the observed output actually exceeds the fitted value. This violates the assumption that the regression provides an upper limit on the producible output. In the Stochastic Frontier Analysis (SFA) literature, this problem has been addressed in various ways. These include adjusting the intercept upwards to cover all data points from above

S. C. Ray (✉)
Department of Economics, University of Connecticut, Storrs, CT, USA
e-mail: subhash.ray@uconn.edu

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_6

(Greene 1980), specifying a one-sided distribution (like the Gamma distribution) for the error term (Richmond 1974), and a full-blown stochastic frontier incorporating a one-sided error term representing inefficiency alongside a two-sided disturbance capturing random variation in the frontier (Aigner et al. 1977).

An alternative to the parametric econometric analysis is the nonparametric approach of Data Envelopment Analysis (DEA) introduced by Charnes et al. (1978). In neoclassical production theory, the intellectual roots of DEA go all the way back to Debreu (1951), Koopmans (1951), Shephard (1953, 1970), Farrell (1957), Farrell and Fieldhouse (1962), Afriat (1972), and Hanoch and Rothschild (1972) among many others. In this strand of the efficiency literature, one makes only a number of general assumptions about the properties of the underlying production technology but leaves the production function unspecified. The observed input-output data are used to solve various mathematical programming problems to measure the technical efficiency of a firm (as in Farrell (1957) or Farrell and Fieldhouse (1962)) or to ascertain whether there exists any production technology satisfying the assumption relative to which the data would be consistent with optimizing behavior by the firms (Afriat 1972; Hanoch and Rothschild 1972; Diewert and Parkan 1983; Varian 1984; Banker and Maindiratta 1988).

The parametric approach of SFA has several advantages over DEA. Being an econometric approach, it readily yields standard errors of the parameter estimates and allows application of standard statistical tests. Moreover, one can derive marginal productivities and various elasticities from the estimated model. However, validity of the entire empirical analysis rests critically on the validity of the functional form specified. If, for example, the true technology is log quadratic (translog) but one estimates a Cobb Douglas production function, specification error may be interpreted as inefficiency. Moreover, the estimated model may violate regularity conditions like nonnegative marginal productivities or negative own-price elasticities of conditional input demand either globally or at individual data points. When the estimated model itself is invalid, little insight can be gained from standard errors of the coefficients.

In DEA, the regularity conditions are imposed on the technology but no functional form is specified. Hence, (weak) monotonicity of the production function or convexity of isoquants is guaranteed to hold. A major weakness of DEA is that it treats all deviations from the frontier as inefficiency and does not readily accommodate random noise. It is possible, however, to generate an empirical distribution function of technical efficiency through bootstrapping and construct a confidence interval.

This chapter presents an overview of the DEA approach to measurement of technical efficiency. Special attention is paid to alternative returns to scale assumptions about the technology. The rest of the chapter is organized as follows. Section 2 introduces the production possibility set (PPS) as the foundation of neoclassical production economics and the Shephard Distance Function as a way to measure the proximity of an observed input-output bundle to the frontier of the PPS. Section 3 details the nonparametric methodology of production analysis and measurement of output- and input-oriented radial technical efficiency under alternative returns to scale assumptions. Section 4 presents the so-called multiplier model form of the DEA linear programming problem and derives the CCR ratio measure of technical efficiency directly from the transformation function for a multiple-output-multiple-input technology. Section 5 looks at the mathematical programming model of Aigner and Chu (1968) for estimating a deterministic parametric frontier production frontier and links it to DEA. Section 6 deals with measurement of scale efficiency and uses the most productive scale size (MPSS) to identify the nature of local returns to scale at different input-output bundles both those which are on the frontier and those which are below the frontier. We also show how to measure scale elasticity in DEA. Section 7 explains non-radial measures of technical efficiency. Section 8 covers Graph efficiency measurement including Graph Hyperbolic efficiency, Directional Distance Function, and Pareto-Koopmans efficiency. Section 9 is the conclusion.

2 Conceptual Foundations

The logical starting point of any discussion of production efficiency is the concept of the PPS. An input vector $x^0 \in R^n_+$ and an output vector $y^0 \in R^m_+$ together constitute a feasible production plan if y^0 can be produced from x^0. The PPS consists of all feasible production plans and is defined as $T = \{(x, y) : y \in R^m_+ \text{ can be produced from } x \in R^n_+\}$.

It is assumed that T is a closed set. In any given context, the size and shape of the set T depend not only on the state of technical knowledge but also on a host of physical, legal, and cultural factors. In the single-output case, it is a common practice to define a production function: $y^* = f(x)$; $\frac{\partial f}{\partial x_i} \geq 0, (i = 1, 2, \ldots, n)$ where y^* is the maximum quantity of output producible from a given input bundle x. The PPS can then be defined as

$$T = \{(x, y) : y \leq f(x),\ y \in R_+, x \in R^n_+\}. \tag{1}$$

Note that while y^* is the maximum output producible from the input bundle x, the actual output (y) can be less than y^* due to inefficiency. For multiple-output-multiple-input production, one uses the transformation function: $F(x,y)=0;\ \frac{\partial F}{\partial x_i}\le 0,\ (i=1,2,\ldots,n);\ \frac{\partial F}{\partial y_j}\ge 0,\ (j=1,2,\ldots,m)$. In that case

$$T=\left\{(x,y):F(x,y)\le 0;\ x\in R_+^m, x\in R_+^n\right\}. \tag{2}$$

The definition in (1) is a special case of the more general definition in (2), with $F(x,y)=y-f(x)$. If $F(x^0,y^0)=0,(x^0,y^0)$ is a boundary point of T. Any reduction in all inputs or increase in all outputs will result in a strictly positive value of the transformation function and the new input-output bundle will be infeasible. The *graph* of the technology[1] is the set $G=\{(x,y):F(x,y)=0;\ x\in R_+^m, x\in R_+^n\}$.

If $F(x^0,y^0)<0,(x^0,y^0)$ is an interior point. Similarly, if $F(x^0,y^0)>0,(x^0,y^0)$ is infeasible.

2.1 Distance Functions and Technical Efficiency

Technical efficiency in production lies in producing the maximum output quantity (or a maximal output bundle) from a given bundle of inputs or in using a minimal input bundle to producing a given output quantity (or bundle). The question of efficiency in production was first formally addressed by Debreu (1951) who introduced the concept of the 'Coefficient of Resource Utilization.' It was essentially a macroeconomic concept measuring production efficiency at the economy level. Technical efficiency of any individual producer depends on the location of its actual input-output bundle relative to the boundary of the PPS or the graph of the technology. Shephard (1953) introduced the distance function for the one output production technology (subsequently generalized to the multiple-output-multiple-input case by Shephard [1970] and McFadden [1978])[2] which provides a measure of technical efficiency of an individual production decision-making unit.

The Output Distance Function evaluated at a specific input-output bundle (x^0,y^0) relative to the PPS (T) is $D^y(x^0,y^0)=\min\lambda:\ (x^0,\frac{1}{\lambda}y^0)\in T$. If one represents the PPS by the transformation function as in (2), the

[1]Some writers prefer to call it the *Graph of the transformation function*.

[2]Although the volume was published in 1978, McFadden's paper has been available since 1968.

Output Distance Function becomes $D^y(x^0, y^0) = \min \lambda : F(x^0, \frac{1}{\lambda} y^0) \leq 0$. If $D^y(x^0, y^0) > 1$, the output bundle y^0 is not producible from x^0 so that $(x^0, y^0) \notin T$. On the other hand, if $D^y(x^0, y^0) \leq 1$, $F(x^0, y^0) \leq 0$ and $(x^0, y^0) \in T$. Of course, if $D^y(x^0, y^0) = 1$, $(x^0, y^0) \in G$.

Thus, an alternative way to define the PPS is $T = \{(x, y) : D^y(x, y) \leq 1; x \in R^m_+, x \in R^n_+\}$.

Comparable to the Output Distance Function is the Input Distance Function $D^x(x^0, y^0) = \max \delta : F(\frac{1}{\delta} x^0, y^0) \leq 0$. Thus, yet another alternative definition of the PPS is $T = \{(x, y) : D^x(x, y) \geq 1; x \in R^m_+, x \in R^n_+\}$.

As noted earlier, technical efficiency precludes the possibility of increasing outputs without increasing inputs or reducing inputs without reducing outputs. At this point, it is important to distinguish between *weak* and *strong* efficiency. An input-output bundle $(x, y) \in T$ is weakly efficient in its output orientation if $\alpha > 1 \Rightarrow (x, \alpha y) \notin T$. Similarly, $(x, y) \in T$ is weakly efficient in its input-orientation if $\beta < 1 \Rightarrow (\beta x, y) \notin T$. As is quite apparent, weak efficiency rules out simultaneous increase in *all outputs* leaving the inputs unchanged or simultaneous reduction in *all inputs* without changing outputs. However, it does not rule out potential increase in one or several outputs or reduction in one or several inputs.[3] By contrast, $(x, y) \in T$ is strongly output efficient if $y' \geq y \Rightarrow (x, y') \notin T$. Here the vector inequality $y' \geq y$ means that y' is at least as large as y in every coordinate and is strictly larger in some coordinate(s). Thus, strong technical efficiency in output orientation rules out increasing any individual output without reducing any output or increasing any input. Similarly, $(x, y) \in T$ is strongly input efficient if $x' \leq x \Rightarrow (x', y) \notin T$. Weak efficiency allows the presence of slacks in some (but not all) outputs (or inputs). Strong efficiency, on the other hand, does not allow slacks in any output (or input). An input-output bundle is Pareto efficient if there is no slack in any output *and* any input.

The output-oriented radial technical efficiency of a firm producing output y^0 from the input x^0 is $\tau_y(x^0, y^0) = 1/\varphi^*$, where $\varphi^* = \max \varphi : (x^0, \varphi y^0) \in T$. Clearly, $\tau_y(x^0, y^0) = D^y(x^0, y^0)$. The input-oriented radial technical efficiency is

$$\tau_x(x^0, y^0) = \min \theta : (\theta x^0, y^0) \in T. \tag{3}$$

$$\tau_x(x^0, y^0) = 1/D^y(x^0, y^0).$$

[3] This can happen if $\frac{\partial F}{\partial y_j} = 0$ for some outputs or $\frac{\partial F}{\partial x_i} = 0$ for some inputs.

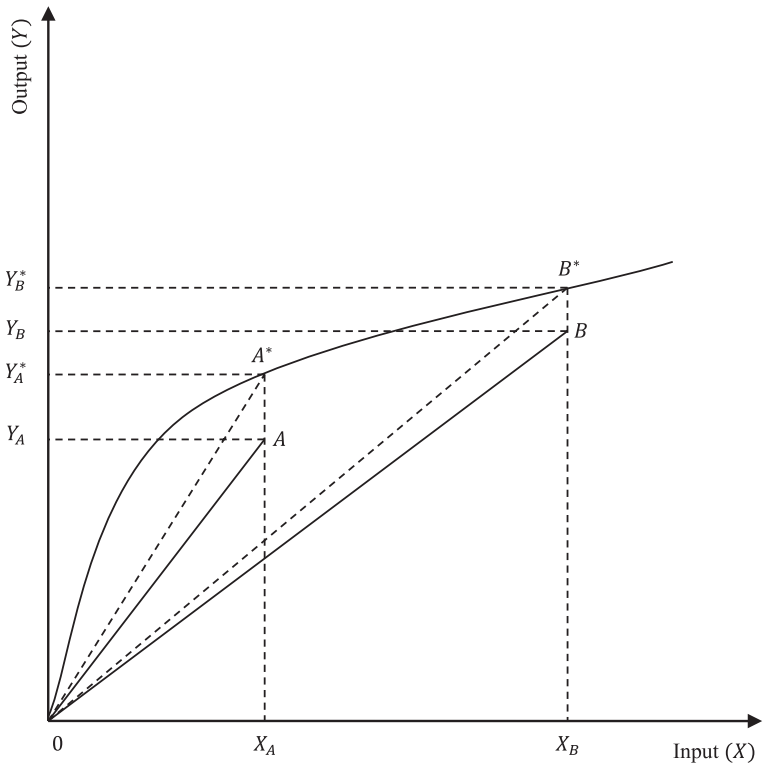

Fig. 1 Output-oriented technical efficiency

Farrell (1957) was the first to formally define and measure technical efficiency in production using input-output data.[4] However, the first mathematical formulation of technical efficiency measurement as a Linear Programming (LP) optimization problem can be found in the appendix of Farrell and Fieldhouse (1962) where is it set up as an output-oriented model.[5]

Measurement of output- and input-oriented technical efficiencies is shown graphically in Figs. 1 and 2. In both figures, the curve OP is the graph of the production function $y = f(x)$ and the points A and B are the input-output quantities (X_A, Y_A) and (X_B, Y_B). In Fig. 1, the points A^* and B^* are their output-oriented efficient projections. Hence,

[4]The computation method in Farrell's paper was extremely burdensome. It involved comparing the input vector per unit of the output used by any unit with convex combinations of all other pairs of such (per unit) input bundles. The technical efficiency of the unit under evaluation was the minimum contraction factor feasible in its unit input bundle compared to any such convex combination. As is apparent from Diagram 2 on page 256 of Farrell (1957), it was an input-oriented measure.

[5]Of course, under the CRS assumption (as shown below) the output- and input-oriented measures of technical efficiency are identical.

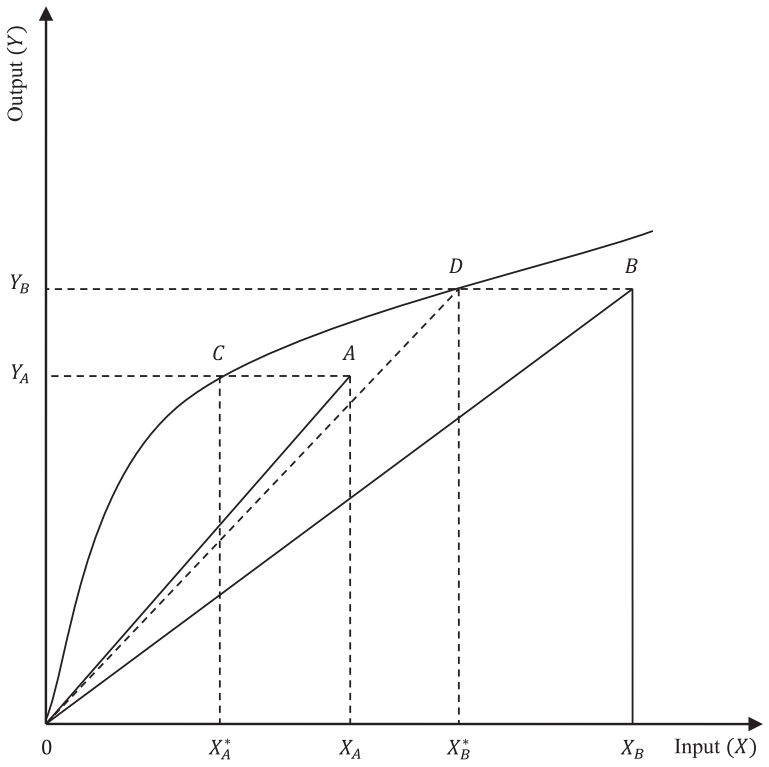

Fig. 2 Input-oriented technical efficiency

$\tau_y(X_A, Y_A) = AX_A/A^*X_A$ and $\tau_y(X_B, Y_B) = BX_B/B^*X_B$. Similarly, the points C and D in Fig. 2 are the input-oriented efficient projections of A and B. Thus, $\tau_x(X_A, Y_A) = CY_A/AY_A$ and $\tau_x(X_B, Y_B) = DY_B/BY_B$.

3 The Nonparametric Methodology

In order to measure technical efficiency in any empirical application, one has to define the PPS by specifying the production function (in the single-output case) or the transformation function (in the multiple-output case). In parametric models, this requires an explicit functional specification like the Cobb Douglas, Constant Elasticity of Substitution (CES), or the Translog production function. Measures of technical efficiency as well as elasticities of substitution between inputs derived from the calibrated model are all contingent upon the validity of the specified functional form. By contrast, in the nonparametric approach of DEA one relies on a number of quite general

assumptions about the underlying technology and employs mathematical programming techniques to construct the frontier of the PPS.

Consider an industry producing m outputs from n inputs. Let (x^j, y^j) be the observed input-output bundle of firm j. Designate the sample of N observations as the set

$$D = \{(x^j, y^j), j = 1, 2, \ldots, N\}.$$

Assumptions

(A1) Every observed input-output bundle is feasible. Thus,

$$(x^j, y^j) \in \Omega \Rightarrow (x^j, y^j) \in T.$$

(A2) The PPS is convex.

$$(x^0, y^0), (x^1, y^1) \in T \Rightarrow \left(\lambda x^0 + (1-\lambda)x^1, \lambda y^0 + (1-\lambda)y^1\right) \in T \quad \text{for all } \lambda \in (0, 1).$$

(A3) Inputs are freely disposable.

$$(x^0, y^0) \in T \wedge x^1 \geq x^0 \Rightarrow (x^1, y^0) \in T.$$

(A4) Outputs are freely disposable.

$$(x^0, y^0) \in T \wedge y^1 \leq y^0 \Rightarrow (x^0, y^1) \in T.$$

One can use the dataset D and the assumptions (A1–A4) to construct the PPS empirically as

$$\hat{T} = \left\{(x, y) : x \geq \sum_{j=1}^{N} \lambda_j x^j; y \leq \sum_{j=1}^{N} \lambda_j y^j; \sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0; \ (j = 1, 2, \ldots, N)\right\}. \quad (4)$$

The intuition behind (4) is quite simple. By virtue of (A1) and (A2), every convex combination $(\bar{x}, \bar{y}) = \left(\sum_{j=1}^{N} \lambda_j x^j, \sum_{j=1}^{N} \lambda_j y^j | \sum_{j=1}^{N} \lambda_j = 1\right)$ is feasible. Further, by (A3), $x \geq \bar{x}$ implies $(x, \bar{y})$ is feasible. Finally by (A4) $y \leq \bar{y}$ implies (x, y) is feasible. The set $\hat{T}$ is often described as the free disposal convex hull of the observed input-output bundles and is the smallest set satisfying assumptions (A1–A4). The frontier of this empirically constructed PPS *envelops* the observed data points most tightly from above. Hence, measuring efficiency using this frontier as the benchmark for evaluation is called DEA.

3.1 Output- and Input-Oriented Radial Technical Efficiency

Banker, Charnes, and Cooper (BCC) (1984) formulated the following LP model to measure the output-oriented technical efficiency of a firm using input $x^0 \in R^n_+$ and producing output $y^0 \in R^m_+$:

$$\begin{aligned} \max \varphi \text{ s.t. } & \sum_{j=1}^{N} \lambda_j y_r^j \geq \varphi y_r^0 \ (r = 1, 2, \ldots, m); \\ & \sum_{j=1}^{N} \lambda_j x_i^j \leq x_i^0 \ (i = 1, 2, \ldots, n); \\ & \sum_{j=1}^{N} \lambda_j = 1; \ \lambda_j \geq 0, \ (j = 1, 2, \ldots, N); \\ & \varphi \text{ unrestricted.} \end{aligned} \tag{5}$$

The solution of (5) yields $\tau_y(x^0, y^0) = 1/\varphi^*$. Even though φ is unrestricted, when (x^0, y^0) is one of the bundles in D (say the bundle of firm k), $\left(\lambda_k = 1, \ \lambda_j = 0 \ (j \neq k), \ \varphi = 1\right)$ is a feasible solution and, in that case, 1 would be a lower bound for φ. But even when (x^0, y^0) is not one of the observed bundles nonnegativity of the λs and the outputs will ensure that φ will never be negative.[6]

The benchmark input-output bundle for (x^0, y^0) is $\left(x^* = \sum_{j=1}^{N} \lambda_j^* x^j, \ y^* = \sum_{j=1}^{N} \lambda_j^* y^j\right)$ constructed from the optimal solution of the problem. For any output r, the difference between the left-hand side and the right-hand side of the relevant output constraint, $s_r^+ = \sum_{j=1}^{N} \lambda_j^* y_r^j - \varphi^* y_r^0$, is the output slack, representing the additional expansion of the output *feasible beyond the common expansion* by the scalar φ^*. Similarly, the input slack, $s_i^- = x_i^0 - \sum_{j=1}^{N} \lambda_j^* x_i^j$, is the potential reduction in input i. The scalar φ^* shows the factor by which *all outputs* can be expanded without requiring any additional input. In fact, some outputs can be expanded further when there are positive output slacks. Similarly, some inputs can even be reduced if there are positive input slacks. The BCC output-oriented model yields a *radial* measure of technical efficiency because it is

[6]However, if any individual input in the bundle x^0 is smaller than the smallest value of the corresponding input across all observations in the dataset D, (3) will not have a feasible solution.

the inverse of the radial output expansion factor φ^* and does not incorporate the output slacks.

In the single-output case, the optimal value of the objective function in the output-oriented DEA problem (φ^*) yields an *estimate* of the maximum output producible from the input bundle x^0 as $\hat{f}(x^0) = \varphi^* y_0$. The true maximum output that can be produced from x^0 may actually be considerably higher than $\varphi^* y_0$. But we cannot infer that on the basis of the observed input-output bundles without making additional assumptions about the technology. However, *it cannot be any smaller than* $\varphi^* y_0$ if the assumptions (A1–A4) hold. In that sense, it is the most conservative estimate of the frontier output and, hence, $\tau_y(x^0, y_0) = 1/\varphi^*$ is an upper bound on the output orient technical efficiency of the firm.[7]

The corresponding input-oriented technical efficiency of the firm using input x^0 and producing output y^0 may be evaluated as $\tau_x(x^0, y^0) = \theta^*$, where

$$\begin{aligned}\theta^* = \min \theta \text{ s.t. } & \sum_{j=1}^{N} \lambda_j y_r^j \geq y_r^0 \ (r = 1, 2, \ldots, m); \\ & \sum_{j=1}^{N} \lambda_j x_i^j \leq \theta x_i^0 \ (i = 1, 2, \ldots, n); \\ & \sum_{j=1}^{N} \lambda_j = 1; \ \lambda_j \geq 0 \ (j = 1, 2, \ldots, N); \\ & \theta \text{ unrestricted.}\end{aligned}$$

Again, it is obvious that $0 < \theta^* \leq 1$. It should be noted that the benchmark input-output bundle ($x^* = \sum_{j=1}^{N} \lambda_j^* x^j$, $y^* = \sum_{j=1}^{N} \lambda_j^* y^j$) on the frontier for the input-oriented DEA problem will, in general, be different from what was obtained for the output-oriented problem.

3.2 Constant Returns to Scale

If one assumes constant returns to scale (CRS), we get the additional assumption

(A5) $(x, y) \in T$ hence $(kx, ky) \in T \quad$ for all $k \geq 0$.

An implication of the CRS assumption is that in the single-output case the production function $y^* = f(x)$ is homogenous of degree 1. That is,

[7]We use superscripts for vectors and subscripts for scalars.

$f(kx) = kf(x)$. In the multiple-output case, CRS implies that the transformation function is homogeneous of degree 0. That is $F(kx, ky) = F(x, y)$. This ensures that if $F(x^0, y^0) \leq 0$, *then* $F(kx^0, ky^0) \leq 0$. Hence, if (x^0, y^0) is feasible, so is (kx^0, ky^0).

Under the additional assumption of CRS, the empirically constructed PPS is

$$\hat{T}_C = \left\{ (x, y) : x \geq \sum_{j=1}^{N} \lambda_j x^j;\ y \leq \sum_{j=1}^{N} \lambda_j y^j;\ \lambda_j \geq 0;\ \ (j = 1, 2, \ldots, N) \right\}. \quad (6)$$

To understand why the constraint $\sum_{j=1}^{N} \lambda_j = 1$ is no longer included consider the following. We have seen above that under (A1–A2), $(\bar{x}, \bar{y}) = \left(\sum_{j=1}^{N} \lambda_j x^j, \sum_{j=1}^{N} \lambda_j y^j | \sum_{j=1}^{N} \lambda_j = 1 \right)$ is feasible. But now, with the added assumption of CRS, $(k\bar{x}, k\bar{y}) = \left(\sum_{j=1}^{N} \lambda_j x^j, \sum_{j=1}^{N} \lambda_j y^j | \sum_{j=1}^{N} \lambda_j = k; k \geq 0. \right)$ is feasible for any $k \geq 0$. But nonnegativity of k is automatically satisfied by the nonnegativity constraints on the $\lambda_j s$ and no additional constraint on the sum of the λs is needed.

The CCR output-oriented CRS DEA LP model is

$$\begin{aligned} \varphi_C^* = \max \varphi_C \text{ s.t. } & \sum_{j=1}^{N} \lambda_j y_r^j \geq \varphi_C y_r^0 \ (r = 1, 2, \ldots, m); \\ & \sum_{j=1}^{N} \lambda_j x_i^j \leq x_i^0 \ (i = 1, 2, \ldots, n); \\ & \lambda_j \geq 0, \ (j = 1, 2, \ldots, N); \\ & \varphi \text{ unrestricted.} \end{aligned} \quad (7)$$

The CRS output-oriented radial technical efficiency is $\tau_y^C(x^0, y^0) = 1/\varphi_C^*$.

The corresponding CRS input-oriented model is

$$\begin{aligned} \theta_C^* = \min \theta \text{ s.t. } & \sum_{j=1}^{N} \lambda_j y_r^j \geq y_r^0 \ (r = 1, 2, \ldots, m); \\ & \sum_{j=1}^{N} \lambda_j x_i^j \leq \theta x_i^0 \quad (i = 1, 2, \ldots, n); \\ & \lambda_j \geq 0 \ (j = 1, 2, \ldots, N); \\ & \theta \text{ unrestricted.} \end{aligned} \quad (8)$$

The CRS input-oriented technical efficiency is $\tau_x^C(x^0, y^0) = \theta_C^*$.

To prove that under CRS the input- and output-oriented radial measures of technical efficiency are identical, we first rewrite the objective function in (7) as $\psi_C^* = \min 1/\varphi_C$ and then divide the constraints by φ_C to rewrite the problem as $\min 1/\varphi_C$ s.t. $\sum_{j=1}^{N} \frac{\lambda_j}{\varphi_C} y_r^j \geq y_r^0$ $(r = 1, 2, \ldots, m)$; $\sum_{j=1}^{N} \frac{\lambda_j}{\varphi_C} x_i^j \leq \frac{1}{\varphi_C} x_i^0$ $(i = 1, 2, \ldots, n)$; $\lambda_j \geq 0$ $(j = 1, 2, \ldots, N)$; φ_C unrestricted. Now define $\psi = 1/\varphi_C$ and $\mu_j = \lambda_j/\varphi_C$. As explained before, even though in principle φ_C is unrestricted in sign, nonnegativity of outputs ensures that it will never be negative. Hence μ_j will also be nonnegative for each observation *j*. Hence, the CRS output-oriented problem can be reduced to

$$\begin{aligned} \min \psi \text{ s.t. } & \sum_{j=1}^{N} \mu_j y_r^j \geq y_r^0 (r = 1, 2, \ldots, m); \\ & \sum_{j=1}^{N} \mu_j x_i^j \leq \psi x_i^0 (i = 1, 2, \ldots, n); \\ & \mu_j \geq 0, (j = 1, 2, \ldots, N); \\ & \psi \text{ unrestricted.} \end{aligned} \tag{9}$$

The problem in (9) is exactly the same as the input-oriented problem in (8). Hence, ψ^* in (9) equals θ_C^* in (8) and $1/\varphi_C^*$ from (7). This proves that $\tau_y^C(x^0, y^0) = \tau_x^C(x^0, y^0)$.

4 The CCR Ratio and the Multiplier Model

In their seminal paper introducing DEA, Charnes et al. (1978) defined technical efficiency as

$$\begin{aligned} h(x^0, y^0) = \max & \sum_{r=1}^{m} u_r y_r^0 / \sum_{i=1}^{n} v_i x_i^0 \\ \text{s.t. } & \sum_{r=1}^{m} u_r y_r^j / \sum_{i=1}^{n} v_i x_i^j \leq 1 \ (j = 1, 2, \ldots, N); \\ & u_r, v_i \geq 0 \ (r = 1, 2, \ldots, m; \ i = 1, 2, \ldots, n). \end{aligned} \tag{10}$$

In the OR/MS literature, the numerator $\sum_{r=1}^{m} u_r y_r^0$ is described as the *virtual output* and the denominator $\sum_{i=1}^{n} v_i x_i^0$ as the *virtual input* of the unit under evaluation. In that sense, $h(x^0, y^0)$ is a measure of total factor

productivity rather than of technical efficiency. Because $(x^0, y^0) \in D$ the constraint in (10) ensures that $h(x^0, y^0)$ cannot exceed unity. Further, because the objective function and the constraints are all homogeneous of degree 0 in u and v, we can normalize the aggregation weights as $\sum_{i=1}^{n} v_i x_i^0 = 1$. The linear fractional functional programming problem in (10) can then be equivalently written as the following LP problem:

$$\begin{aligned} h(x^0, y^0) = \max & \sum_{r=1}^{m} u_r y_r^0 \\ \text{s.t.} & \sum_{r=1}^{m} u_r y_r^j - \sum_{i=1}^{n} v_i x_i^j \leq 0 \ (j = 1, 2, \ldots, N); \\ & \sum_{i=1}^{n} v_i x_i^0 = 1; \ u_r, v_i \geq 0 \ (r = 1, 2, \ldots, m; \ i = 1, 2, \ldots, n). \end{aligned} \tag{11}$$

The problem in (11) is the linear programming dual of the optimization problem for the CRS input-oriented technical efficiency in (8). Hence, by standard duality results, the CCR ratio $h(x^0, y^0)$ in (11) equals $\tau_x^C(x^0, y^0)$ from (8) and under CRS also equals $\tau_y^C(x^0, y^0)$.

Ray (2019) provides a derivation of the CCR ratio directly from the transformation function. Consider the input-output bundle (x^0, y^0) and assume that $F(x^0, y^0) < 0$ so that (x^0, y^0) is an inefficient bundle. Next, consider the Shephard Distance Function $D^y(x^0, y^0) = \min\beta : (x^0, y^0/\beta) \in T \Leftrightarrow F(x^0, y^0/\beta) = 0$. It is the inverse of the largest scalar δ such that $F(x^0, \delta y^0) = 0$ and is the same as the output-oriented Farrell efficiency. Clearly, for $(x^0, y^0), \delta > 1$ and $\beta < 1$.

Focus now on the efficient input-output bundle $(x^0, y_0^*) = (x^0, \delta y^0)$ lying on the graph of the technology. Thus $F(x^0, y_0^*) = 0$. Now, due to homogeneity of degree 0,

$$F(x^0, y_0^*) = \sum_i \left(\frac{\partial F}{\partial x_i}\right)_{x^0, y_0^*} x_{i0} + \sum_r \left(\frac{\partial F}{\partial y_r}\right)_{x^0, y_0^*} y_{r0}^* = 0. \tag{12}$$

Define $F_i^0 \equiv \left(\frac{\partial F}{\partial x_i}\right)_{x^0, y_0^*} (i = 1, 2, \ldots, n)$ and $F_r^0 \equiv \left(\frac{\partial F}{\partial y_r}\right)_{x^0, y_0^*} (r = 1, 2, \ldots, m)$. Then (12) becomes

$$\sum_i F_i^0 x_{i0} + \sum_r F_r^0 y_{r0}^* = 0. \tag{13}$$

Because $y^*_{r0} = \delta y_{r0}$, (13) leads to $\delta \sum_r F^0_r y_{r0} = -\sum_i F^0_i x_{i0}$. Thus, $\beta = 1/\delta = -\sum_r F^0_r y_{r0} / \sum_i F^0_i x_{i0}$. Define $u_r \equiv F^0_r$ and $v_i \equiv -F^0_i$ to derive $\beta = 1/\delta = \sum_r u_r y_{r0} / \sum_i v_i x_{i0}$. This shows that the CCR ratio in (10) is the same as the Shephard Distance Function or the Farrell measure of technical efficiency. Finally, $\sum_{r=1}^m u_r y_r - \sum_{i=1}^n v_i x_i = 0$ is a supporting hyperplane to the PPS at $(x^0, \delta^* y^0)$ and due to convexity $\sum_{r=1}^m u_r y^j_r - \sum_{i=1}^n v_i x^j_i \le 0$ for all $(x^j, y^j) \in D$. The multipliers (u and v) are the vectors of shadow prices of outputs and inputs. In that sense, the CCR ratio is the 'shadow return on outlay' and the inequality constraints are comparable to the 'no positive profit' condition in a competitive market under CRS.

The dual LP problem for the output-oriented VRS DEA (also known as the BCC) model is

$$\begin{aligned} &\min\ v_0 + \sum_{i=1}^{n} v_i x^0_i \\ &\text{s.t. } v_0 + \sum_{i=1}^{n} v_i x^j_i - \sum_{r=1}^{m} u_r y^j_r \ge 0\ (j = 1, 2, \ldots, N); \\ &\qquad \sum_{i=1}^{n} u_r y^0_r = 1; u_r, v_i \ge 0\ (r = 1, 2, \ldots, m;\ i = 1, 2, \ldots, n). \end{aligned} \tag{14}$$

In the next section, we show how the BCC-DEA problem relates to the deterministic parametric frontier model of Aigner and Chu (1968).

5 A Deterministic Parametric Frontier as a Special DEA Problem

As noted earlier, in the single-output case the observed input-output data must satisfy the inequality $y \le f(x)$. It is convenient, therefore, to write the deterministic production frontier[8] as

$$y = f(x)\, e^{-u},\ u \ge 0. \tag{15}$$

Aigner and Chu (1968) specified the Cobb-Douglas form of the frontier production function $f(x) = A x_1^{\beta_1} x_2^{\beta_2} \ldots x_k^{\beta_k}$. Then (15) can

[8]In the parametric frontier production function literature, (39) is described as a deterministic frontier. It needs to be made clear that what is deterministic above is the frontier output $y^* = f(x)$ but the actual output y is stochastic because e^{-u}, the inefficiency component, is still random.

be written in logarithmic form for each individual observation j as $u_j = \beta_0 + \beta_1 \ln x_1^j + \beta_2 \ln x_2^j + \cdots + \beta_k \ln x_k^j - \ln\, y_j \geq 0;\ \beta_0 = \ln(A)$. Given the one-sided distribution of u, the usual OLS estimation procedure does not work in this case. Aigner and Chu proposed minimizing either $\sum_{j=1}^N u_j^2$ or $\sum_{j=1}^N u_j$ subject to the constraints that each estimated residual u_j is nonnegative. Note that minimizing $\sum_{j=1}^N u_j$ is equivalent to minimizing $\frac{1}{N}\sum_{j=1}^N u_j = \bar{u} = \beta_0 + \beta_1 \overline{\ln x_1} + \beta_2 \overline{\ln x_2} + \cdots + \beta_k \overline{\ln x_k} - \overline{\ln y}$. Further, $\overline{\ln y}$ is a constant and plays no role in the minimization problem. Hence, the Aigner-Chu problem can be formulated as

$$\begin{aligned} &\min\ \beta_0 + \beta_1 \overline{\ln x_1} + \beta_2 \overline{\ln x_2} + \cdots + \beta_k \overline{\ln x_k} \\ &\text{s.t. } \beta_0 + \beta_1 \ln x_1^j + \beta_2 \ln x_2^j + \cdots + \beta_k \ln x_k^j \geq \ln\, y_j; \\ &\qquad \beta_i \geq 0\ (i = 1, 2, \ldots, k);\ \beta_0 \text{ unrestricted.} \end{aligned} \tag{16}$$

The LP dual to (16) is

$$\begin{aligned} &\max \sum_{j=1}^N \lambda_j \ln y_j \\ &\text{s.t. } \sum_{j=1}^N \lambda_j \ln x_i^j \leq \overline{\ln x_i}\ (i = 1, 2, \ldots, m), \\ &\qquad \sum_{j=1}^N \lambda_j = 1;\ \lambda_j \geq 0\ (j = 1, 2, \ldots, N). \end{aligned} \tag{17}$$

Next, define $\varphi \equiv \sum_{j=1}^N \lambda_j \ln y_j / \overline{\ln y}$. Because $\lambda_j = 1/N$ for all j is a feasible solution for (17), $\sum_{j=1}^N \lambda_j \ln y_j \geq \varphi \overline{\ln y}$. Hence, (17) is equivalent to

$$\begin{aligned} \max \varphi \text{ s.t. } &\sum_{j=1}^N \lambda_j \ln y_j \geq \varphi \overline{\ln y}; \\ &\sum_{j=1}^N \lambda_j \ln x_i^j \leq \overline{\ln x_i}\ (i = 1, 2, \ldots, m); \\ &\sum_{j=1}^N \lambda_j = 1;\ \lambda_j \geq 0\ (j = 1, 2, \ldots, N). \end{aligned}$$

Thus, the Aigner-Chu model actually solves the BCC-DEA output-oriented problem evaluating the efficiency of the geometric mean of the input-output bundle in the (log) input-output space.[9]

6 Scale Efficiency

The concept of scale efficiency is best explained for a 1-output 1-input technology. Consider a firm with the input-output pair (x_0, y_0). At the observed point, the average productivity of the firm is $\mathrm{AP}_0 = y_0/x_0$. If the firm is inefficient, $y_0 < f(x_0)$ and its output can be increased to $f(x_0)$. As a result, its average productivity would increase to $\mathrm{AP}(x_0) = f(x_0)/x_0$. Thus, eliminating technical inefficiency would automatically raise average productivity. However, once the firm has been projected on to the frontier, no further increase in average productivity is possible without changing the input level. All points on the frontier are technically efficient. However, unless CRS holds globally across the frontier, average productivity will vary across different points on the frontier. We may define average productivity along the frontier as $\mathrm{AP}(x) = \frac{f(x)}{x}$. Frisch (1965) defined the input level, x^*, where $\mathrm{AP}(x)$ reaches a maximum as the *technically optimal production scale* (TOPS). A measure of scale efficiency of the firm would then be $\mathrm{SE}(x_0) = \mathrm{AP}(x_0)/\mathrm{AP}(x^*)$. It is well known from standard microeconomics that the first-order condition for a maximum of $\mathrm{AP}(x)$ is $f(x)/x = f'(x)$ or $f(x) = xf'(x)$. Hence, at the input level x^*, $\mathrm{AP}(x^*) = f'(x^*)$. Define the constant $\kappa = f'(x^*)$ and a counterfactual CRS production function $r(x) = \kappa x$. Two things may be emphasized. First, by the definition of the TOPS, for every input level x

$$\mathrm{AP}(x) = f(x)/x \leq \mathrm{AP}(x^*) = f(x^*)/x^* = f'(x^*) = \kappa \quad \text{hence } f(x) \leq \kappa x = r(x).$$

Second, $\mathrm{SE}(x_0) = \mathrm{AP}(x_0)/\mathrm{AP}(x^*) = f(x_0)/(\kappa x_0) = f(x_0)/r(x_0) \leq 1$, which can also be written as $\mathrm{SE}(x_0) = \frac{y_0/r(x_0)}{y_0/f(x_0)}$. Now, $y^* = f(x)$ is the true VRS frontier and $y_0/f(x_0) = \tau_y(x_0, y_0)$ is a measure of the output-oriented VRS technical efficiency. On the other hand, $y^{**} = r(x)$ is an artificial CRS frontier and $y_0/r(x_0) = \tau_y^C(x_0, y_0)$ is a measure of the output-oriented CRS technical efficiency. Hence, a simple measure of scale efficiency is $\mathrm{SE}(x_0) = \tau_y^C(x_0, y_0)/\tau_y(x_0, y_0)$.

[9]For an MLE interpretation of the Aigner-Chu model, see Schmidt (1976). Banker (1993) shows that DEA provides an MLE of a deterministic frontier. For details of econometric estimation of a parametric stochastic frontier see Kumbhakar and Lovell (2000).

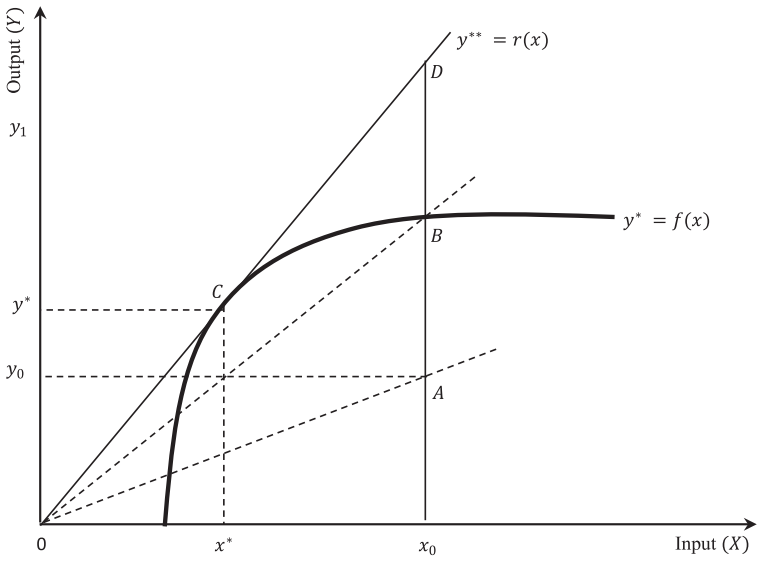

Fig. 3 Scale efficiency

In Fig. 3, the point A shows the input-output bundle (x_0, y_0) and the point B on the production function $y^* = f(x)$ is the output-oriented technically efficient projection. The most productive input scale is x^* and $\text{AP}(x^*) = f(x^*)/x^* = Cx^*/Ox^* = Dx_0/Ox_0$. Also, the tangent to the production function at the point C can be treated as a counterfactual CRS production function $y^{**} = r(x) = \kappa x; \kappa \equiv f'(x^*)$. Thus,

$$\begin{aligned}\text{SE}(x_0) &= \frac{f(x_0)/x_0}{f(x^*)/x^*} = Bx_0/Dx_0 \\ &= \frac{Ax_0/Dx_0}{Ax_0/Bx_0} = f(x_0)/r(x_0) \\ &= \frac{y_0/r(x_0)}{y_0/f(x_0)} = D_C^y(x_0, y_0)/D^y(x_0, y_0).\end{aligned}$$

6.1 Ray Average Productivity and Returns to Scale

The concept of average productivity is unequivocal only in a single-output single-input case. When multiple inputs are used for production (as is almost universally the case in real life), one can measure partial average productivities but to obtain a single measure of total factor productivity one

must aggregate the individual inputs into a scalar. In the multiple-output case, the individual outputs also need to be aggregated. One way to avoid such aggregation is to consider only proportional variations in all inputs and the consequent proportional variation in all outputs. Consider a technically efficient pair of input-output bundles (x^0, y^0) satisfying $F(x^0, y^0) = 0$. Now consider another input bundle $x^1 = \beta x^0$. Thus, the two input bundles have the same input-proportions and differ only in scale. Next consider an output bundle y satisfying $F(x^1, y) = 0$. There will be many output bundles y satisfying $F(x^1, y) = 0$. Out of them we select the one which is proportional to y^0. Denote this as $y^1 = \alpha y^0$. Now compare the bundles (x^0, y^0) and $(x^1, y^1) = (\beta x^0, \alpha y^0)$. Note that because outputs increase (decrease) with inputs, $\beta > 1$ implies $\alpha > 1$. If we treat x^0 as a single unit of a composite input and y^0 as a single unit of a composite output then (x^1, y^1) corresponds to β units of the composite input and α units of the composite output. In terms of the composite input and output, the *ray average productivity* at (x^0, y^0) is 1 and at (x^1, y^1) it is $y^0 \in R^m_+$ Now, if $\alpha > \beta$, we conclude that ray average productivity has increased and increasing returns to scale holds locally at (x^0, y^0). Similarly, locally DRS holds when $\frac{\alpha}{\beta} < 1$. Of particular interest is a bundle for which $\frac{\alpha}{\beta} = 1$ showing locally CRS.

Note that between any two points (x, y) and $(x + \mathrm{d}x, y + \mathrm{d}y)$ both on the frontier

$$\begin{aligned} \mathrm{d}F &= \sum_{i=1}^{n} F_i \mathrm{d}x_i + \sum_{j=1}^{m} F_j \mathrm{d}y_j = 0 \\ \text{hence } \sum_{i=1}^{n} F_i x_i \mathrm{d}x_i / x_i &+ \sum_{j=1}^{m} F_j y_j \mathrm{d}y_j / y_j = 0. \end{aligned} \tag{18}$$

But when the input bundles are proportional, $\mathrm{d}x_i / x_i$ equals some constant q_1 for every input i. Similarly, proportionality of the output bundles implies $\mathrm{d}y_j / y_j$ equals a constant q_2 for every output j. Hence, from (18)

$$\begin{aligned} q_1 \sum_{i=1}^{n} F_i x_i + q_2 \sum_{j=1}^{m} F_j y_j &= 0 \\ \text{hence } q_2 / q_1 = -\sum_{i=1}^{n} F_i x_i / \sum_{j=1}^{m} F_j y_j. \end{aligned} \tag{19}$$

Now, $x + \mathrm{d}x = (1 + q_1)x$ and $y + \mathrm{d}y = (1 + q_2)y$. Define $\beta = 1 + q_1$ and $\alpha = 1 + q_2$. Then $q_2 > q_1$ implies $\alpha > \beta$ and locally IRS holds at (x, y).

Similarly, $q_2 < q_1$ implies locally DRS. Finally, $q_2 = q_1$ implies locally CRS. Starrett (1977) defined the *degree of increasing returns* as $\text{DIR} = \frac{q_2}{q_1} - 1$.

6.2 Most Productive Scale Size and Local Returns to Scale

Banker (1984) defined the *most productive scale size* (MPSS) in the context of multiple-output-multiple-input production as a generalization of Frisch's concept of the TOPS. According to his definition, *an input-output bundle* (x^0, y^0) *is an MPSS if for all nonnegative real numbers* (α, β) *such that* $(\beta x^0, \alpha y^0)$ *is feasible*, $\alpha/\beta \leq 1$. In other words,(x^0, y^0) is an MPSS if and only if there is no other input-output bundle proportional to it with a higher ray average productivity. Obviously, an inefficient bundle cannot be an MPSS because both its output-oriented projection $(x^0, \varphi^* y^0)$ and its input-oriented projection $(\theta^* x^0, y^0)$ will have a higher ray average productivity.

6.3 Identifying Local Returns to Scale at an Efficient Bundle

In the DEA literature, there are three alternative ways to identify the local returns to scale properties of an input-output bundle that lies on the frontier of the PPS (a) Banker's Primal approach, (b) a dual approach to due to BCC, and (c) a nesting approach due to Färe et al. (1985). Of course, all of them lead to the same conclusion.

Banker's Primal Approach

Assume that the bundle (x^0, y^0) is efficient under VRS. The following theorem in Banker (1984) identifies whether or not it is an MPSS.

Theorem: *An input-output bundle* (x^0, y^0) *is an MPSS if and only if the optimal value of the objective function of a CCR-DEA model equals unity for this input-output combination.*

Proof See Banker (1984, p. 40).

This theorem only determines if the bundle (x^0, y^0) is an MPSS. If it is not, all we know is that locally CRS does not hold at this point on the frontier. That does not, by itself, pick between IRS and DRS. However, the following corollaries from the theorem can be used to identify local RTS properties by looking at the optimal solution of the CCR-DEA problem:

1. If $k = \sum_{j=1}^{N} \lambda_j^* = 1$, (x^0, y^0) is an MPSS and CRS holds locally.
2. If $k = \sum_{j=1}^{N} \lambda_j^* < 1$, IRS holds locally at (x^0, y^0).
3. If $k = \sum_{j=1}^{N} \lambda_j^* > 1$, DRS holds locally at (x^0, y^0).

Note that the corollaries (1–3) hold both for the output-oriented problem in (7) and (8) the input-oriented problem in (8). One potential problem with this approach is that there may be alternative optimal solutions to the CCR-DEA LP problem with the optimal values of the λs adding up to greater than 1 in one and less than 1 in another. To resolve this ambiguity, one should modify the corollaries above and require that the respective conditions for IRS and DRS should hold at all optimal solutions of the CCR-DEA problem.

To implement this, one would first solve either the output- or the input-oriented CRS DEA problem (7) or (8) and obtain the optimal value. Suppose that one selected the output-oriented model and obtained φ^* from (7). Next one would use φ^* as a parameter to solve the following problem

$$\begin{aligned} \max k &= \sum_{j=1}^{N} \lambda_j \\ \text{s.t. } &\sum_{j=1}^{N} \lambda_j y_r^j \geq \varphi^* y_r^0 \ (r = 1, 2, \ldots, m); \\ &\sum_{j=1}^{N} \lambda_j x_i^j \leq x_i^0 \ (i = 1, 2, \ldots, n); \\ &\lambda_j \geq 0 \ (j = 1, 2, \ldots, N). \end{aligned} \tag{20}$$

One would conclude that the condition in corollary 2 holds in all optimal solutions of (7) if the optimal k^* in (20) is less than 1. To check for the condition in corollary 3, one would set up (20) as a minimization problem and conclude that DRS holds at (x^0, y^0) if the minimum of k is greater than 1.

The BCC Dual Approach

Banker, Charnes, and Cooper (BCC) (1984) offer an alternative method of identifying local returns to scale from the following dual of the output-oriented VRS DEA problem shown in (14) above.

BCC have shown that

i. CRS holds at (x^0, y^0) if at the optimal solution of (14) v_0 is zero;
ii. IRS holds at (x^0, y^0) if at the optimal solution of (14) v_0 is < 0;
iii. DRS holds at (x^0, y^0) if at the optimal solution of (14) v_0 is > 0.

If there are alternative optimal solutions for (14), conditions (i)–(iii) must hold at all optimal solutions.

A simple proof of (i)–(iii) follows from Ray (2019). Because (x^0, y^0) is efficient by assumption, at the optimal solution of (14) $v_0^* + \sum_{i=1}^n v_i^* x_i^0 = 1 = \sum_{i=1}^n u_r^* y_r^0$. hence $v_0^* + \sum_{i=1}^n v_i^* x_i^0 - \sum_{i=1}^n u_r^* y_r^0 = 0$. But, as defined above, $u_r \equiv F_r^0$ and $v_i \equiv -F_i^0$ and in the present context (19) can also be written as $q_2/q_1 = -\sum_{i=1}^n F_i x_i / \sum_{j=1}^m F_j y_j = \sum_{i=1}^n v_i^* x_i^0 / \sum_{j=1}^m u_r^* y_r^0$ or

$$q_2/q_1 - 1 = \left(\sum_{i=1}^{n} v_i^* x_i^0 - \sum_{j=1}^{m} u_r^* y_r^0 \right) / \sum_{j=1}^{m} u_r^* y_r^0. \tag{21}$$

Because the denominator in (21) is always positive, the sign of the ratio is determined by the sign of the numerator. Specifically, when IRS holds, $q_2/q_1 - 1 > 0 \Rightarrow \sum_{i=1}^n v_i^* x_i^0 - \sum_{j=1}^m u_r^* y_r^0 = -v_0^* > 0 \Rightarrow v_0^* < 0$. By the same logic, DRS implies $q_2/q_1 - 1 > 0$ and $v_0^* > 0$. Finally, for CRS $q_1 = q_2$ and $v_0^* = 0$. Of course, as in the case of Banker's approach, multiple optimal solutions pose a problem and the conditions (ii) and (iii) have to be appropriately modified.

A Nesting Approach

Färe et al. (1985) consider a technology that lies in between CRS and the VRS technologies. They call it a non-increasing returns to scale (NIRS) technology. Under the assumption of NIRS $(x^0, y^0) \in T \Rightarrow (kx^0, ky^0) \in T;\ 0 \leq k \leq 1$. Thus, any feasible input-output bundle remains feasible if it is scaled downwards but not necessarily feasible if scaled upwards.

The DEA estimate of an NIRS PPS is

$$\hat{T}^N = \left\{ (x, y) : x \geq \sum_{j=1}^{N} \lambda_j x^j;\ y \leq \sum_{j=1}^{N} \lambda_j y^j; \right. \\ \left. \sum_{j=1}^{N} \lambda_j \leq 1;\ \lambda_j \geq 0\ (j = 1, 2, \ldots, N) \right\} \tag{22}$$

The three different sets in (4), (6), and (22) are nested so that $\hat{T} \subset \hat{T}^N \subset \hat{T}^C$. Because the VRS PPS is the most and the CRS PPS is the least restrictive, the measured technical efficiency will be the highest under VRS and lowest under CRS. The frontiers of the CRS and NIRS production possibility sets coincide in the region of IRS. Similarly, the VRS and NIRS frontiers are identical in the DRS region. Therefore, when IRS holds at (x^0, y^0), in an input-oriented model $\theta_*^C = \theta_*^N < \theta_*^V$, where the superscripts C, N, and V refer to CRS, NIRS, and VRS. Similarly, $\theta_*^C < \theta_*^N = \theta_*^V$ implies DRS. Of course, in the case of CRS, all three estimates of technical efficiency equal unity. Note that in this nesting approach the possibility of multiple optimal solutions does not pose any problem because the objective function value does not differ across alternative optimal solutions.

6.4 Returns to Scale Properties of an Inefficient Input-Output Bundle

As has been noted earlier, returns to scale is a meaningful concept only when the input-output bundle is efficient and is a point on the frontier of the PPS. In the foregoing discussion, it was assumed that (x^0, y^0) is technically efficient. When that is not the case, one must first project it on to the frontier and only then can examine the local RTS properties at the efficient projection. This, however, creates an ambiguity because there would be two alternative projections—the input-oriented $(\theta^* x^0, y^0)$ and the output-oriented $(x^0, \varphi^* y^0)$. The nature of local RTS can be different at the two different points.

At this point, it would be helpful to define for any given input-output pair the 2-dimensional conditional PPS:

$$T(x^0, y^0) = \{(\alpha, \beta) : (\beta x^0, \alpha y^0) \in T;\ \alpha, \beta \geq 0\}. \tag{23}$$

In terms of the transformation function $T(x^0, y^0) = \{(\alpha, \beta) : F(\beta x^0, \alpha y^0) \leq 0;\ \alpha, \beta \geq 0\}$. Similarly, the conditional graph of the technology is

$$G(x^0, y^0) = \{(\alpha, \beta) : F(\beta x^0, \alpha y^0) = 0;\ \alpha, \beta \geq 0\}. \tag{24}$$

One can think of the graph in (24) as the *ray production function*

$$\alpha = g(\beta). \tag{25}$$

The MPSS for (x^0, y^0) corresponds to the highest ray average productivity along (25). The following lemma from Ray (2010) shows that when the PPS is convex, IRS holds at all scales smaller than the smallest MPSS. Similarly, DRS holds at all scales larger than the largest MPSS.[10]

Lemma: *For any convex productivity possibility set T, if there exist nonnegative scalars* α *and* β *such that* $\alpha > \beta > 1$, *and both* $(\tilde{x}, \tilde{y})$ *and* $(\beta\tilde{x}, \alpha\tilde{y}) \in G$, *then* $\gamma > \delta$ *for every* γ *and* δ *such that* $1 < \delta < \beta$ *and* $(\delta\tilde{x}, \gamma\tilde{y}) \in G$.

Proof: Because $(\tilde{x}, \tilde{y})$ and $(\beta\tilde{x}, \alpha\tilde{y})$ are both feasible, by convexity of T, for every $\lambda \in (0, 1)$, $((\lambda + (1-\lambda)\beta)\tilde{x}, (\lambda + (1-\lambda)\alpha)\tilde{y})$ is also feasible. Now select λ such that $\lambda + (1-\lambda)\beta = \delta$. Further, define $\mu = \lambda + (1-\lambda)\alpha$. Using these notations, $(\delta\tilde{x}, \mu\tilde{y}) \in T$. But, because $(\delta\tilde{x}, \gamma\tilde{y}) \in G$, $\gamma \geq \mu$. However, because $\alpha > \beta$, $\mu > \delta$. Hence, $\gamma > \delta$.

An implication of this lemma is that, when the PPS is convex, if the technology exhibits locally diminishing returns to scale at any smaller input scale, it cannot exhibit increasing returns at a bigger input scale. This is easily understood in the single-input single-output case. When both x and y are scalars, average productivity at $(\tilde{x}, \tilde{y})$ is $\tilde{y}/\tilde{x}$ and at $(\beta\tilde{x}, \alpha\tilde{y})$ is $(\alpha/\beta)\tilde{y}/\tilde{x}$. Thus, when $\alpha > \beta$, average productivity has increased. The above lemma implies that for every input level x in between $\tilde{x}$ and $\beta\tilde{x}$, average productivity is greater than $\tilde{y}/\tilde{x}$. Thus, average productivity could not first decline and then increase as the input level increased from $\tilde{x}$ and $\beta\tilde{x}$.

Two results follow immediately. First, locally increasing returns to scale holds at every input-output bundle $(x, y) \in G$ that is smaller than the smallest MPSS. Second, locally diminishing returns to scale holds at every input-output bundle $(x, y) \in G$ that is greater than the largest MPSS. To see this, let $x = bx^*$ and $y = ay^*$, where (x^*, y^*) is the smallest MPSS for the given input and output mix. Because (x, y) is not an MPSS, $a/b < 1$. Further, assume that $b < 1$. Define $\beta = 1/b$ and $\alpha = 1/a$. Then $(x^*, y^*) = (\beta x, \alpha y)$ and $\alpha/\beta > 1$. Because ray average productivity is higher at a larger input scale, by virtue of the lemma, locally increasing returns to scale holds at (x, y). Next assume that $b > 1$. Again, because (x, y) is not an MPSS, $a/b < 1$. That is ray average productivity has fallen as the input scale is increased from x^* to $x = bx^*$. Then, by virtue of the lemma, ray average product could not be any higher than a/b at a slightly greater input scale, $\bar{\bar{x}} = (1 + \varepsilon)x$. But, because (x, y) is not an MPSS, ray average product cannot remain constant as the

[10] For a different proof see Banker and Thrall (1992).

input scale is slightly increased. Hence, ray average product must fall as the input scale is slightly increased from x. Thus, locally diminishing returns to scale holds at every $(x, y) \in G$, when x is larger than the largest MPSS.

6.5 Finding the MPSS

One can solve the following optimization problem proposed by Cooper et al. (1996) to find the MPSS for the input-output bundle (x^0, y^0):

$$\begin{aligned} \max \alpha/\beta \text{ s.t. } & \sum_{j=1}^{N} \lambda_j y^j \geq \alpha y^0; \\ & \sum_{j=1}^{N} \lambda_j x^j \leq \beta x^0; \\ & \sum_{j=1}^{N} \lambda_j = 1; \\ & \alpha, \beta, \lambda_j \geq 0, (j = 1, 2, \ldots, N). \end{aligned} \tag{26}$$

Even though the objective function is nonlinear, it can be easily transformed into a linear programming problem by defining $t = 1/\beta$ and $\mu_j = t\lambda_j$ ($j = 1, 2, \ldots, N$). Nonnegativity of β and λ_js ensures that t and μ_js are also nonnegative. Problem (26) can, therefore, be reformulated as the following linear programming problem:

$$\begin{aligned} \max \rho \text{ s.t. } & \sum_{j=1}^{N} \mu_j y^j \geq \rho y^0; \\ & \sum_{j=1}^{N} \mu_j x^j \leq \beta x^0; \\ & \sum_{j=1}^{N} \mu_j = t; \\ & t, \mu_j \geq 0 \ (j = 1, 2, \ldots, N). \end{aligned} \tag{27}$$

From the optimal solution of this problem, we can derive $\beta^* = 1/t^*$ and $\alpha^* = \rho^*/t^*$. One can then infer the nature of returns to scale from these values of α^* and β^*. It may be pointed out here that because the only

restriction on t is nonnegativity, (27) is simply the output-oriented CCR-DEA problem and $1/\rho^*$ is the same as the output-oriented CRS technical efficiency $\tau_y^C(x^0, y^0)$. Thus, (27) is in reality a minor modification of Banker (1984).

Because (x^0, y^0) is a feasible input-output bundle, $(\alpha = \beta = \rho = 1)$ is a feasible solution for this problem. Hence, the optimal value ρ^* is always greater than or equal to 1.When $\rho^* = \alpha^*/\beta^*$ exceeds unity, we know that (x^0, y^0) *is not an MPSS*. But, we can also conclude that $(\beta^* x^0, \alpha^* y^0)$ *is an MPSS.* When the bundle (x^0, y^0) is not itself an MPSS, $\rho^* > 1$ so that $\alpha^* > \beta^*$.

We may invoke the lemma above to determine the local RTS properties at (x^0, y^0) by comparing it with its MPSS. Note that in the conditional PPS (23), (x^0, y^0) corresponds to $(\alpha, \beta) = (1, 1)$. Let the MPSS be designated as $(\beta^* x^0, \alpha^* y^0) = (x^*, y^*)$. If the MPSS is unique, there are five different possibilities:

i. $1 < \beta^* < \alpha^*$;
ii. $\beta^* < \alpha^* < 1$;
iii. $\beta^* < 1 < \alpha^*$;
iv. $\beta^* = 1 < \alpha^*$;
v. $\beta^* < 1 = \alpha^*$.

In case (i), $x^0 < \beta^* x^0 = x^*$ and $y^0 < \alpha^* y^0 = y^*$. Thus, (x^0, y^0) lies toward the southwest of the MPSS. Both input- and output-oriented projections of the bundle (x^0, y^0) fall in the region of IRS. In this case, the unit is conclusively too small relative to its MPSS. Similarly, in case (ii) $x^0 > \beta^* x^0 = x^*$ and $y^0 > \alpha^* y^0 = y^*$ and both input- and output-oriented projections fall in the region of DRS. The implication is that the unit is too large. In case (iii), the RTS characterization depends on the direction of projection. Because $\beta^* < 1$, $x^* < x^0$, and the input scale is bigger than the MPSS. The output-oriented projection falls in the region of DRS. At the same time, because $1 < \alpha^*$, $y^0 < y^*$ the output scale is smaller than the MPSS and the input-oriented projection falls in the region of IRS. In case (iv) $x^0 = x^*$ but $y^0 < y^*$. This time the input scale corresponds to the MPSS but the output scale is too small. Eliminating output-oriented technical inefficiency will automatically project the observed bundle on to its MPSS. Similarly in case (v) $y^0 = y^*$ but $x^0 > x^*$. The input-oriented efficient projection is the MPSS.

Figure 4 shows how the local RTS properties of an input-output bundle can be determined by comparing it to its MPSS. In this diagram, the scale of the input bundle x^0 is measured as β along the horizontal axis and the scale of the output bundle y^0 is measured as α up the vertical axis.

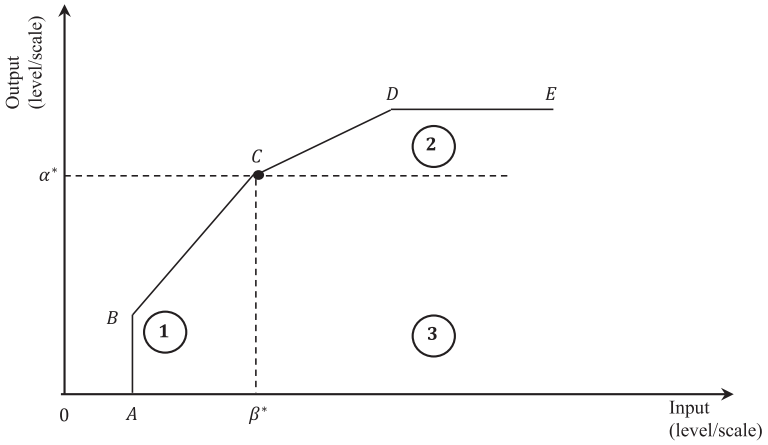

Fig. 4 MPSS and RTS regions

The piecewise connected line *ABCDE* shows the pairs (α, β) for which the corresponding input-output bundle $(\beta x^0, \alpha y^0)$ is on the frontier of the PPS. Two points of special interest are $(\alpha, \beta) = (1, 1)$ and the point C, $(\alpha, \beta) = (\alpha^*, \beta^*)$. The former is the location of the observed bundle (x^0, y^0) and the latter is its MPSS. The local RTS properties of (x^0, y^0) depends on where $(\alpha, \beta) = (1, 1)$ is located relative to C in this diagram. This is true even when (1, 1) lies on the *ABCDE* line and is technically efficient. If (1, 1) lies in area (1) to the southwest of C, both input- and output-oriented projections will be smaller than the MPSS and IRS holds unequivocally at (x^0, y^0). If it lies in area (2) toward the northeast of C both projections will be larger than the MPSS and DRS holds at (x^0, y^0). By contrast, area (3) is an inclusive region. Here the output-oriented projection is larger than the MPSS implying DRS but the input-oriented projection is smaller than the MPSS implying IRS. The unit is too small judged by its output scale but is too large when judged by the input scale. Case (iv) corresponds to points on the vertical line $C\beta^*$ while points on the horizontal line through C below the frontier correspond to case (v).

The Case of Multiple MPSS

Next we consider the possibility of multiple MPSS. This is depicted graphically in Fig. 5. Here both C_1 and C_2 are MPSS and so are their convex combinations lying on the line segment connecting them. At C_1, (α_1^*, β_1^*) is the smallest MPSS. Similarly, (α_2^*, β_2^*) at C_2 is the largest MPSS. It is obvious that when the problem in (27) has a unique optimal solution (in particular, t^* is unique), there cannot be multiple MPSS. For multiple optimal

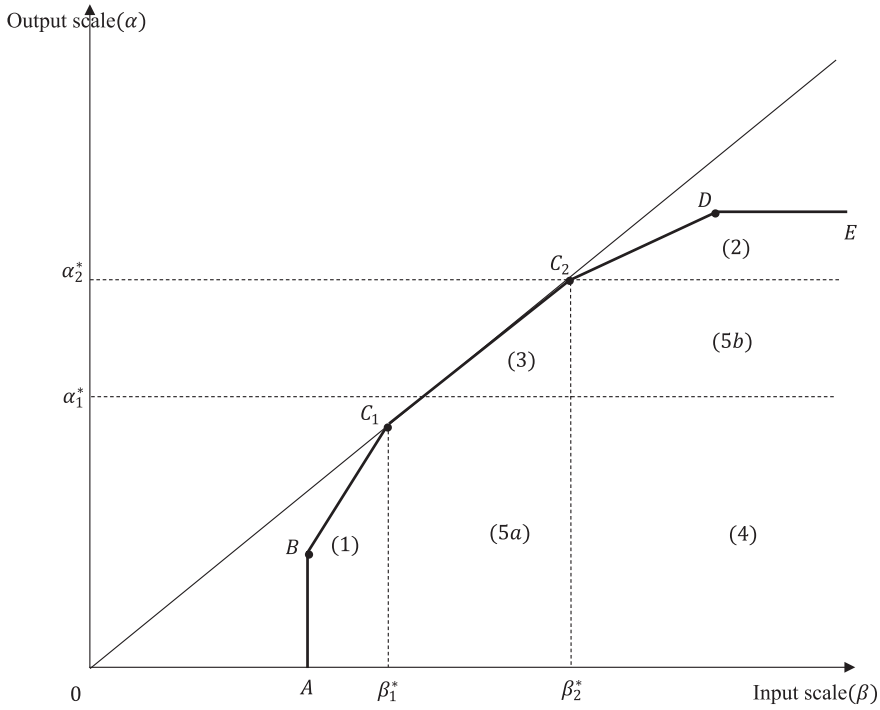

Fig. 5 Multiple MPSS and the regions of the increasing, decreasing and the ambiguous returns to scale

solutions, the largest $t^* = \sum_j \mu_j^*$ across all optimal solutions of (27) corresponds to the smallest MPSS, β_1^*. Similarly, β_2^* corresponds to the smallest $t^* = \sum_j \mu_j^*$ at an optimal solution.

Since across all optimal solutions the value of the objective function is the same (ρ^*), $\beta_1^* = 1/t_1^*$, where

$$\begin{aligned} t_1^* = \max & \sum_j \mu_j \\ \text{s.t.} & \sum_j \mu_j x^j \le x^0; \\ & \sum_j \mu_j y^j \ge \rho^* y^0; \\ & \mu_j \ge 0 \ (j = 1, 2, \ldots, N). \end{aligned} \tag{28}$$

Similarly, $\beta_2^* = 1/t_2^*$, where

$$
\begin{aligned}
t_2^* = \min & \sum_j \mu_j \\
\text{s.t.} & \sum_j \mu_j x^j \leq x^0; \\
& \sum_j \mu_j y^j \geq \rho^* y^0; \\
& \mu_j \geq 0 \ (j = 1, 2, \ldots, N).
\end{aligned} \tag{29}
$$

Once β_1^* and β_2^* have been determined from (28) and (29), the corresponding values of α are readily obtained as $\alpha_1^* = \rho^* \beta_1^*$ and $\alpha_2^* = \rho^* \beta_2^*$.

As shown in Fig. 5, the set of output-input scales (α, β) for which the input-output bundles $(\beta x^0, \alpha y^0)$ are feasible can be partitioned into six different regions defined below:

i. In region (1), toward the southwest of the smallest MPSS (C_1), $(\beta < \beta_1^*; \alpha < \alpha_1^*)$. When (x^0, y^0) falls in this region, $1 < \beta_1^* < \alpha_1^*$. Hence, increasing returns to scale holds unambiguously.
ii. In region (2), to the northeast of the largest MPSS (C_2), $(\beta_2^* < \beta; \alpha_2^* < \alpha)$. If (x^0, y^0) falls in this region, $\beta_1^* < \alpha_1^* < 1$. Diminishing returns to scale holds unambiguously in this region.
iii. In region (3), $\beta_1^* < \beta < \beta_2^*$ while $\alpha_1^* < \alpha < \alpha_2^*$. Points in this region lie between the smallest and the largest MPSS. It is interesting to note, that even if the point (α=1, β=1) is not technically efficient and lies below the C_1 C_2 line, both the input- and the output-oriented projection of the inefficient bundle will fall in the region of CRS. Thus, there is no scale inefficiency in this region even though there may be technical inefficiency.
iv. In region (4), $\beta_2^* < \beta$; $\alpha < \alpha_1^*$. When the actual input-output bundle lies here, $\beta_2^* < 1 < \alpha_1^*$. The input bundle x^0 is larger than the largest MPSS hence the output-oriented projection falls in the area of diminishing returns. At the same time, the actual output bundle is smaller than the smallest MPSS. Hence, increasing returns to scale holds at the input-oriented projection. Thus, returns to scale cannot be unambiguously defined at the actual input-output bundle.
v. In region (5a), $\beta_1^* < \beta < \beta_2^*$ but $\alpha < \alpha_1^*$. When the actual input-output bundle lies here, y^0 is smaller than the smallest MPSS and the input-oriented projection falls in the area of increasing returns. At the same time, the actual input bundle lies between the smallest and the largest MPSS. Hence, CRS holds at the output-oriented projection. Here also the returns to scale characterization depends on the orientation.

vi. In region (5b), $\beta_2^* < \beta$ while $\alpha_1^* < \alpha < \alpha_2^*$. When the actual input-output bundle lies here, x^0 is larger than the largest MPSS. Hence the output-oriented projection falls in the area of diminishing returns. At the same time, the actual output bundle lies between the smallest and the largest MPSS. Hence, CRS holds at the input-oriented projection. Here the input bundle is too large. But the actual output bundle, if produced from the technically efficient input bundle would correspond to an MPSS.

6.6 Scale Elasticity

The foregoing analysis provides only qualitative information about the returns to scale characteristics of an efficient point located on the frontier or the output (or input) oriented projection of an inefficient onto the frontier. By contrast, the *scale elasticity* measured at a point on the frontier provides a quantitative measure of the proportionate change in the output in the single-output case (and equi-proportionate change in all outputs, in the multi-output case) relative to an equi-proportionate change in all inputs.

The textbook definition of scale elasticity in the 1-output multi-input case with the production function $y = f(x)$ is $\varepsilon_0 = \frac{\partial \ln f(tx)}{\partial \ln t}|_{t=1} = \sum_i \frac{\partial \ln f_i}{\partial \ln x_i}$. Frisch (1965) define this as the *passus coefficient*. In the Anglo-Saxon literature is also defines the *function coefficient* (e.g., Ferguson 1969, p. 79). Returns to scale for multi-output technologies has been considered by Hanoch (1970), Starrett (1977), and Panzar and Willig (1977) among others. Among the various papers on nonparametric measures of scale elasticity with multiple outputs using DEA, the most significant ones are by Banker et al. (1984), Banker and Thrall (1992), Førsund (1996), and Førsund and Hjalmarsson (2004).[11]

Consider the input-output bundle (x^0, y^0) where $x^0 \in R_+^n$ and $y^0 \in R_+^m$. Further suppose that (x^0, y^0) is on the frontier of the PPS so that $F(x^0, y^0) = 0$. Now consider the bundle $(\beta x^0, \alpha y^0)$ which also is on the frontier so that

$$F(\beta x^0, \alpha y^0) = 0. \tag{30}$$

Førsund (1996)[12] defined the scale elasticity at (x^0, y^0) as $\varepsilon = \frac{d \ln \alpha}{d \ln \beta}|_{\alpha=1,\beta=1} = \frac{d\alpha}{d\beta} \cdot \frac{\beta}{\alpha}|_{\alpha=1,\beta=1} = \frac{d\alpha}{d\beta}$. Differentiating (30) with respect

[11]For a discussion of scale efficiency in the context of Global Returns to Scale (GRS) see Podinovski (2017).
[12]Unlike here, Førsund (1996) used μ for the output scaling factor and β for the input scaling factor.

to the input scaling factor β, $\sum_i \frac{\partial F(\beta x^0, \alpha y^0)}{\partial x_i} x_i^0 + \sum_i \frac{\partial F(\beta x^0, \alpha y^0)}{\partial y_j} y_j^0 \frac{d\alpha}{d\beta} = 0$, and evaluating at $\alpha = \beta = 1$, we obtain $\varepsilon = \frac{d \ln \alpha}{d \ln \beta} = \frac{d\alpha}{d\beta} = -\sum_i \frac{\partial F(x^0, y^0)}{\partial x_i} x_i^0 / \sum_i \frac{\partial F(x^0, y^0)}{\partial y_j} y_j^0$. Using the previous definitions, $v_i \equiv -\frac{\partial F(x^0, y^0)}{\partial x_i}$ and $u_j \equiv \frac{\partial F(x^0, y^0)}{\partial y_j}$, we can write $\varepsilon = \frac{d \ln \alpha}{d \ln \beta} = \frac{d\alpha}{d\beta} = \sum_i v_i x_i^0 / \sum_i u_j y_j^0$. Comparing this with the radial VRS output-oriented DEA model shown in (14) at its optimal solution, we can see that $\sum_j u_j y_j^0 = 1$. Further, because (x^0, y^0) is efficient by assumption, the optimal value of the output-oriented primal problem in (5) equals 1 and hence by standard duality results, the optimal value in (14) also equals 1: $v_0 + \sum_{i=1}^n v_i x_i^0 = 1$. This implies that $\sum_{i=1}^n v_i x_i^0 = 1 - v_0$ and, therefore,

$$\varepsilon = \sum\nolimits_i v_i x_i^0 / \sum\nolimits_i u_j y_j^0 = 1 - v_0. \tag{31}$$

Equation (21) of Førsund and Hjalmarsson (2004, p. 1030) obtains the scale elasticity measure as $\varepsilon(x^0, y^0) = 1 - E_2 v_0$ where E_2 is their notation for output-oriented efficiency. Because (x^0, y^0) is efficient by assumption, E_2 equals unity and their formula reduces to (31) above.

But what if (x^0, y^0) is not on the frontier? We must then first radially project it to the frontier. Taking the output-oriented projection, the point under consideration will be (x^0, y_*^0) where $y_*^0 = \varphi^* y^0$ and φ^* is the maximum output scaling factor. Now we start from $F(\beta x^0, \alpha y_*^0) = 0$. Proceeding as before we end up with $\varepsilon = -\sum_i \frac{\partial F(x^0, y_*^0)}{\partial x_i} x_i^0 / \sum_i \frac{\partial F(x^0, y_*^0)}{\partial y_j} y_{*j}^0 = \sum_i v_i x_i^0 / \sum_i u_j y_{*j}^0 = \sum_i v_i x_i^0 / (\varphi^* \sum_i u_j y_j^0)$. As in (14) and (5), the optimal solution is $\sum_j u_j y_j^0 = 1$ and the optimal value of the objective function is $v_0 + \sum_i v_i x_i^0 = \varphi^*$. Hence,

$$\varepsilon(x^0, \varphi^* y^0) = (\varphi^* - v_0)/\varphi^* = 1 - v_0/\varphi^*. \tag{32}$$

In empirical applications, a potential problem is that the even though the optimal value of the objective function in the VRS output-oriented dual problem will be unique, as recognized before, there may be multiple optimal solutions. Differing values of v_0 across the alternative solutions yield different measures of scale elasticity both in (32) and in (31). Let $v_0^{\min}$ be the minimum and $v_0^{\max}$ the maximum value of v_0 across all optimal solutions of (14).[13] Then the corresponding maximum and minimum values of scale elasticity in (32) are

[13]See Banker and Thrall (1992) section 4 for the relevant procedure for finding the minimum and the maximum values of v_0.

$$\varepsilon^{\max}(x^0, \varphi^* y^0) = 1 - v_0^{\min}/\varphi^*;$$
$$\varepsilon^{\min}(x^0, \varphi^* y^0) = 1 - v_0^{\max}/\varphi^*. \tag{33}$$

One obtains the corresponding maximum and the minimum values for (31) by setting φ^* equal to 1 in (33).

6.7 Global Returns to Scale

All of the foregoing discussion on local returns to scale characterization of a bundle on the frontier (or its efficient projection) if it is an interior point rests critically on the assumption that the PPS T is convex. As already noted, an implication of the lemma in Sect. 4 is that when locally increasing returns to scale is detected, an input-output bundle must increase in scale in order to attain its MPSS. The opposite is true for locally diminishing returns to scale. Such information can become valuable for deciding on proposal to merge several firms or to break up a large firm. As shown by Podinovski (2004a, b), once the convexity assumption is relaxed, one needs to distinguish between local returns to scale and GRS. A simple example of non-convexity can be found in the free disposal hull (FDH) technology which is obtained by relaxing convexity but retaining the assumption of free disposability of inputs and outputs.[14] The frontier of the FDH production possibility set looks like a step function with flat segments followed by jumps. Average productivity declines long the flat segment, followed by a sudden increase at the jump point. For non-convex technologies, the maximum ray average productivity for a given input-output bundle may be attained at multiple scales but not at any intermediate scale between them. Podinovski (2004a) defines each of these scales as a scale reference unity (SRU) and provides the following classification of GRS property of any input-output bundle on the frontier:

The input-output bundle (x^0, y^0) exhibits

[14]The Free Disposal Hull was introduced by Deprins et al. (1984). For a detailed discussion of FDH analysis, see Ray (2004, chapter 6). See also Kerstens and Vanden Eeckaut (1999).

a. Global constant returns to scale (G-CRS) if and only if it is an MPSS;
b. Global diminishing returns to scale (G-DRS) if it is bigger than all of its SRUs;
c. Global increasing returns to scale (G-IRS) if and only if it is smaller than all of its SRUs;
d. Globally sub-constant returns to scale if it is smaller than some its SRUs but bigger than some of its SRUs.

7 Non-radial Measures of Efficiency

A problem with the radial models both input- and output-oriented is that when slacks are present in the input or output constraints at the optimal solution of the relevant DEA problem, the extent of inefficiency is underestimated. In an output-oriented model, for example, a positive slack in any output constraint represents the amount by which the concerned output can be further expanded beyond the common expansion factor (φ^*). In an extreme example, if there is no room to increase one specific output without increasing inputs even if the other outputs can be doubled, the radial efficiency measure of the firm under evaluation will still be 100%.[15]

One way out of this paradoxical situation is to consider a non-radial measure of technical efficiency that reflects the full potential for increasing in every output even though not equi-proportionately. Färe and Lovell (1978) introduced what they called a Russell measure of technical efficiency that rules out the presence of output slacks in an output-oriented model and input slacks in an input-oriented model.[16]

7.1 Output-Oriented Russell Measure

The output-oriented Russell measure of technical efficiency is the inverse of the maximum average expansion factor across the individual outputs and is measured as $\mathrm{RM}_y(x^0, y^0) = 1/\rho_y(x^0, y^0)$, where

[15]In fact, CCR (1979) included a small penalty (ε) for slacks in their DEA models. However, because ε is assumed to be an arbitrarily small (non-Archimedean) number, it cannot be incorporated in any practical application.

[16]Input slacks may still be present at the optimal solution in a non-radial output-oriented model and output slacks in a non-radial input-oriented model.

$$\begin{aligned}\rho_y(x^0, y^0) = \max \tfrac{1}{m} & \sum_{r=1}^{m} \varphi_r \\ \text{s.t.} & \sum_{j=1}^{N} \lambda_j y_r^j \geq \varphi_r y_r^0 \ (r = 1, 2, \ldots, m); \\ & \sum_{j=1}^{N} \lambda_j x_i^j \leq x_i^0 \ (i = 1, 2, \ldots, n); \\ & \sum_{j=1}^{N} \lambda_j = 1; \ \varphi_r \geq 1 \ (r = 1, 2, \ldots, m); \\ & \lambda_j \geq 0 \ (j = 1, 2, \ldots, N).\end{aligned} \tag{34}$$

It can be seen that the radial output-oriented model can be recovered from this non-radial model by imposing the additional restriction $\varphi_r = \varphi$ for each input r $(r = 1, 2, \ldots, m)$. This, the BCC output-oriented measure under VRS (or the CCR measure under CRS) is a special case of the Russell measure and $\tau_y(x^0, y^0) \geq \mathrm{RM}_y(x^0, y^0)$.

7.2 Non-radial Russell Input Efficiency

Analogous to (34) is the input-oriented Russell efficiency measure is the arithmetic mean of the input specific contraction factors (θ_i) across all inputs,

$$\begin{aligned}\mathrm{RM}_x(x^0, y^0) = \rho_x(x^0, y^0) = \min \tfrac{1}{n} & \sum_{i=1}^{n} \theta_i \\ \text{s.t.} & \sum_{j=1}^{N} \lambda_j y_r^j \geq y_r^0 \ (r = 1, 2, \ldots, m); \\ & \sum_{j=1}^{N} \lambda_j x_i^j \leq \theta_i x_i^0 \ (i = 1, 2, \ldots, n); \\ & \sum_{j=1}^{N} \lambda_j = 1; \ \theta_i \leq 1 \ (i = 1, 2, \ldots, n); \\ & \lambda_j \geq 0 \ (j = 1, 2, \ldots, N).\end{aligned}$$

Again, the radial input-oriented technical efficiency is a restricted version of the Russell input-oriented measure with $\theta_i = \theta$ for all i and $\tau_x(x^0, y^0) \geq \mathrm{RM}_x(x^0, y^0)$.

8 Graph Efficiency Measures

The alternative measures of efficiency considered so far are either output-oriented or input-oriented. In an output-oriented model, an inefficient unit is projected on to the frontier by the maximum proportional expansion of all of its outputs, but reducing inputs is not an objective. Similarly, in an input-oriented model, the objective is only to scale down all inputs proportionately as much as possible. In measuring graph efficiency, one seeks to achieve some reduction in inputs side by side with expansion of outputs. The problem in this case would be that unlike in an output-oriented or an input-oriented model, the direction of projection on to the frontier is arbitrary. Two popular measures of graph efficiency are the ones based on the Graph Hyperbolic Distance Function due to Färe et al. (1985) and the Directional Distance Function introduced by Chambers et al. (1996).

8.1 Graph Hyperbolic Efficiency

For the Graph Hyperbolic efficiency measure, one selects the point on the frontier that lies on a rectangular hyperbola through the observed input-output bundle. In a single-output single-input case, both the actual point (x_0, y_0) and the efficient point (x^*, y^*) satisfy the equation $xy = k$. This is shown in Fig. 6 where the point A represents the observed input-output bundle (x_0, y_0) and the point B on the production frontier is the efficient projection (x^*, y^*). The level of efficiency is $\frac{1}{\delta^*} = \frac{Ox^*}{Ox_0} = \frac{Oy_0}{Oy^*}$.

To obtain the efficient projection in the multiple-output case, one needs to solve the problem:

$$\begin{aligned} \max \delta \text{ s.t. } & \sum_{j=1}^{N} \lambda_j y_r^j \geq \delta y_r^0 \ (r = 1, 2, \ldots, m); \\ & \sum_{j=1}^{N} \lambda_j x_i^j \leq \tfrac{1}{\delta} x_i^0 / \delta \ (i = 1, 2, \ldots, n); \\ & \sum_{j=1}^{N} \lambda_j = 1; \ \lambda_j \geq 0 \ (j = 1, 2, \ldots, N); \\ & \delta \text{ unrestricted.} \end{aligned} \tag{35}$$

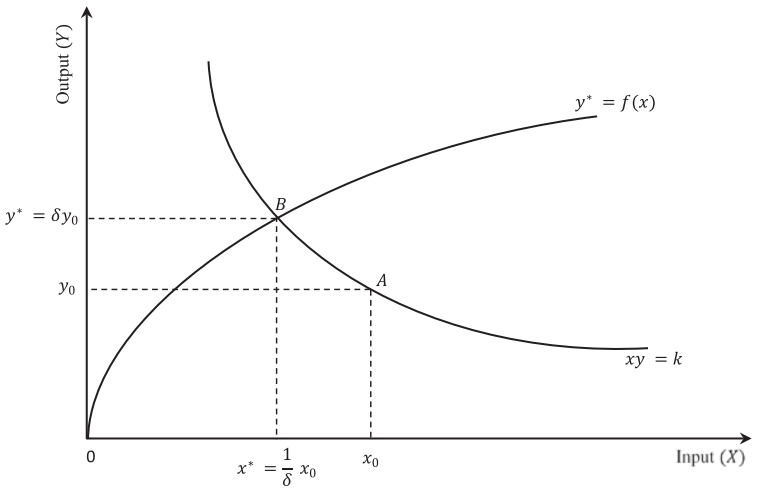

Fig. 6 Graph hyperbolic distance function

The Graph Hyperbolic measure of efficiency is $\tau_{\text{GH}}(x^0, y^0) = 1/\delta^*$. The input constraints in (35) are nonlinear. However, if one assumes CRS, one can define $\mu_j = \delta\lambda_j$ and $\varphi = \delta^2$ to rewrite the model as

$$\begin{aligned} \max \varphi \text{ s.t. } & \sum_{j=1}^{N} \mu_j y_r^j \geq \varphi y_r^0 \ (r = 1, 2, \ldots, m); \\ & \sum_{j=1}^{N} \mu_j x_i^j \leq x_i^0 \ (i = 1, 2, \ldots, n); \\ & \mu_j \geq 0 \ (j = 1, 2, \ldots, N); \\ & \varphi \text{ unrestricted.} \end{aligned}$$

In this case, $\tau_{\text{GH}}^C(x^0, y^0) = 1/\sqrt{\varphi^*}$. In the case of VRS, the problem in (35) remains nonlinear. Färe et al. (1985) linearize the input constraints using the Taylor's series approximation

$$f(\delta) = \tfrac{1}{\delta} \approx f(\delta_0) + f'(\delta_0)(\delta - \delta_0) = (2 - \delta)/\delta_0. \tag{36}$$

Using $\delta_0 = 1$ as the point of approximation, the linearized version of (35) is

$$\begin{aligned}
\max \delta \text{ s.t. } & \sum_{j=1}^{N} \lambda_j y_r^j \geq \delta y_r^0 \ (r = 1, 2, \ldots, m); \\
& \sum_{j=1}^{N} \lambda_j x_i^j + \delta x_i^0 \leq 2x_i^0 \ (i = 1, 2, \ldots, n); \\
& \sum_{j=1}^{N} \lambda_j = 1; \ \lambda_j \geq 0 \ (j = 1, 2, \ldots, N); \\
& \delta \text{ unrestricted.}
\end{aligned} \tag{37}$$

Note that assuming $\delta_0 = 1$ amounts to assuming that the observed point is on the frontier. When this is not the case, the approximation will be rather poor. Ideally, one should use $\delta_0 = 1$ only as the starting point and iterate (36)–(37) until convergence.

8.2 Directional Distance Function

Building upon Luenberger's (1992) *benefit function*, Chambers et al. (1996, 1998) introduced the Directional Distance Function to measure the distance of an observed input-output bundle from the frontier of the PPS in a direction chosen by the analyst. Let $g^x = (g_1^x, g_2^x, \ldots, g_n^x) \in R_+^n$ and $g^y = (g_1^y, g_2^y, \ldots, g_m^y) \in R_+^m$ be two direction subvectors. Then the Directional Distance Function can be defined as $\vec{D}(x^0, y^0; g^x, g^y) = \max\beta : (x^0 - \beta g^x, y^0 + \beta g^y) \in T$. It is clear that one can recover the radial output-oriented model by setting $g^x = 0$ and $g^y = y^0$. In that case,β would equal $(\varphi - 1)$ in (5) or (7), depending on the returns to scale assumption. Another interesting choice of the direction for projection would be $(g^x, g^y) = (x^0, y^0)$. Then $\vec{D}(x^0, y^0; g^x, g^y) = \max\beta : ((1 - \beta)x^0, (1 + \beta)y^0) \in T$ and β is the maximum percentage by which all outputs can be expanded and all inputs can be contracted simultaneously. In Fig. 7, A is the observed bundle (x_0, y_0). The point B $(g^x = -x^0, g^y = y_0)$ defines the direction of movement. The point C on the production frontier shows the maximum feasible movement within the PPS in the direction parallel to OB. In this case, the Directional Distance Function is $\beta = AC/OB = OD/OB$.

For an arbitrary choice of (g^x, g^y), the relevant VRS DEA problem is

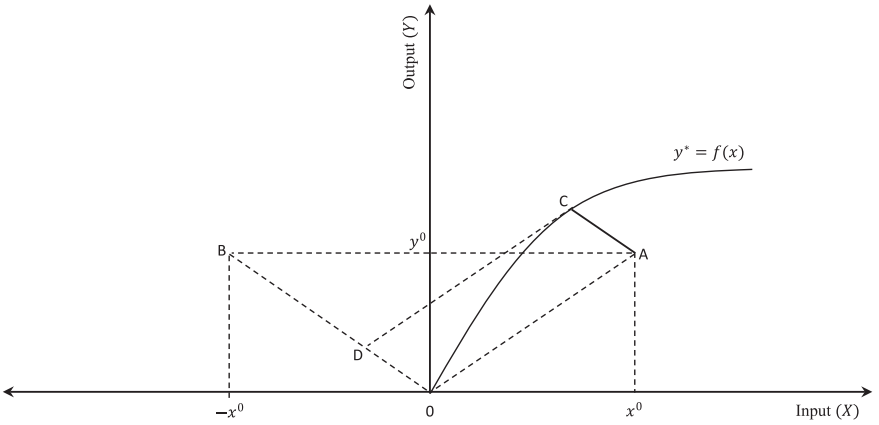

Fig. 7 Directional distance function

$$\begin{aligned}
\max \beta \text{ s.t. } & \sum_{j=1}^{N} \lambda_j y_r^j - \beta g_r^y \geq y_r^0 \ (r = 1, 2, \ldots, m); \\
& \sum_{j=1}^{N} \lambda_j x_i^j + \beta g_i^x \leq x_i^0 \ (i = 1, 2, \ldots, n); \\
& \sum_{j=1}^{N} \lambda_j = 1; \ \lambda_j \geq 0 \ (j = 1, 2, \ldots, N); \\
& \beta \text{ unrestricted.}
\end{aligned} \tag{38}$$

The flexibility of the Directional Distance Function is apparent from the fact that it can be radial (setting $g^x = 0$ or $g^y = 0$), biradial (setting $g^x = x^0$ and $g^y = y^0$), or completely non-radial for arbitrary choice of (g^x, g^y).

Ray (2007) introduced a measure of overall inefficiency as

$$\vartheta(x^0, y^0) = \max (\varphi - \theta) : \ (\theta x^0, \varphi y^0) \in T. \tag{39}$$

In a radial output-oriented model, a measure of technical inefficiency is $\varphi - 1$ where φ is the maximum scaling factor for all outputs. Similarly, the input-oriented technical inefficiency is $1 - \theta$ where θ is the minimum scaling factor for all inputs. In that sense, the overall inefficiency is the sum of output and input inefficiencies.

The DEA LP problem for (39) is

$$\begin{aligned}
\max \varphi - \theta \text{ s.t. } & \sum_{j=1}^{N} \lambda_j y_r^j \geq \varphi y_r^0 \ (r = 1, 2, \ldots, m); \\
& \sum_{j=1}^{N} \lambda_j x_i^j \leq \theta x_i^0 \ (i = 1, 2, \ldots, n); \\
& \sum_{j=1}^{N} \lambda_j = 1; \ \lambda_j \geq 0 \ (j = 1, 2, \ldots, N); \\
& \beta \text{ unrestricted.}
\end{aligned} \tag{40}$$

The dual LP for (40) is

$$\begin{aligned}
\min \pi \text{ s.t.} \pi \geq & \sum_{r=1}^{m} u_r y_r^j - \sum_{v=1}^{n} v_i x^j \ (j = 1, 2, \ldots, N); \\
& \sum_{r=1}^{m} u_r y_0^r = 1; \quad \sum_{i=1}^{n} v_i x_0^i = 1; \\
& u_r \geq 0 \ (r = 1, 2, \ldots, m); \\
& v_i \geq 0 \ (i = 1, 2, \ldots, n); \\
& \pi \text{ unrestricted.}
\end{aligned} \tag{41}$$

Ray (2007) has shown that if the optimal π^* in (41) is positive, then there does not exist any pair of nonnegative shadow price vector (u, v) corresponding to which the bundle (x^0, y^0) would be the profit-maximizing input-output pair.

Further, (40) can be seen to encompass (38). Define $\beta^y = \varphi - 1$ and $\beta^x = 1 - \theta$. Then, (40) becomes

$$\begin{aligned}
\max \beta^y + \beta^x \text{ s.t. } & \sum_{j=1}^{N} \lambda_j y_r^j \geq (1 + \beta^y) y_r^0 \ (r = 1, 2, \ldots, m); \\
& \sum_{j=1}^{N} \lambda_j x_i^j \leq (1 - \beta^x) x_i^0 \ (i = 1, 2, \ldots, n); \\
& \sum_{j=1}^{N} \lambda_j = 1; \ \lambda_j \geq 0 \ (j = 1, 2, \ldots, N); \\
& \beta^x, \ \beta^y \text{ unrestricted.}
\end{aligned}$$

By imposing the restriction $\beta^x = \beta^y$ one gets the problem in (38) except for a scaling factor of the objective function.[17]

8.3 Pareto-Koopmans Measures

A Russell efficient output bundle contains no output slack. Similarly, no input slack can be present in a Russell efficient input bundle. However, to be Pareto-Koopmans efficient an input-output bundle must be simultaneously Russell efficient in both output and input orientations Thus, Pareto-Koopmans (PK) efficiency combines both input- and output-oriented Russell efficiency. There are different variants of this PK efficiency but the most popular of them is the product of the Russell output and input efficiencies.[18] It is called Enhanced Russell Measure by Pastor, Louis, and Sirvent (PLS) (1999), Slack Based Measure by Tone (2001),[19] and simply Pareto-Koopmans efficiency by Ray (2004) and can be measured as

$$\begin{aligned}\tau^{\mathrm{PK}}(x^0, y^0) = \min & \tfrac{1}{n}\sum_i \theta_i / \tfrac{1}{m}\sum_r \varphi_r \\ \text{s.t.} & \sum_j \lambda_j y_{rj} \geq \varphi_r y_{r0} \ (r = 1, 2, \ldots, m); \\ & \sum_j \lambda_j x_{ij} \leq \theta_i x_{i0} \ (i = 1, 2, \ldots, n); \\ & \varphi_r \geq 1 \ (r = 1, 2, \ldots, m); \\ & \theta_i \leq 1 \ (i = 1, 2, \ldots, n); \\ & \sum_j \lambda_j = 1; \\ & \lambda_j \geq 0 \ (j = 1, 2, \ldots, N)\end{aligned} \tag{42}$$

Every input and output constraint in (42) will be strictly binding. Therefore at the optimal projection $x_i^* = \sum_j \lambda_j^* x_{ij} = \theta_i^* x_{i0}$ $(i = 1, 2, \ldots, n)$. Define the total reduction in input i as for each input $s_i^- = x_i^0 - x_i^* \geq 0$.

[17] The model in (38) is further developed in Aparicio et al. (2013).

[18] Portela and Thanassoulis (2005) used the measure $\Pi(\theta_i^{1/n})/\Pi(\varphi_r^{1/m})$ and called it the Geometric Distance Function.

[19] Tone's SBM appeared in 2001 in EJOR but makes no reference to the PLS Enhanced Russell measure introduced in the same journal in 1999 and the two are virtually identical.

This leads to $\theta_i^* = x_i^*/x_{i0} = 1 - s_i^-/x_{i0}$. Similarly by defining $s_r^+ = y_r^* - y_{ro}$, we can derive $\varphi_r^* = x_i^*/x_{i0} = 1 + s_r^+/y_{r0}\,(r = 1, 2, \ldots, m)$. Hence the objective function in (42) becomes $\frac{1}{n}\sum_i \theta_i / \frac{1}{m}\sum_r \varphi_r = (1 - \frac{1}{n}\sum_i s_i^-/x_{i0})/(1 + \frac{1}{m}\sum_r s_{ir}^+/y_{r0})$, which is the Slack Based Measure.[20]

Both PLS and Tone use the expression in (42) for the objective function and resort to a normalization to convert the linear fractional functional programming problem into an LP following Charnes and Cooper (1968). Ray (2004), Ray and Jeon (2009), and Ray and Ghose (2014) on the other hand, use a linear approximation of the objective function at $(\theta_i = 1, \varphi_r = 1)\,(i = 1, 2, \ldots, n;\ r = 1, 2, \ldots, m)$ to get $\frac{1}{n}\sum_i \theta_i / \frac{1}{m}\sum_r \varphi_r \approx 2 + \frac{1}{n}\sum_i \theta_i - \frac{1}{m}\sum_r \varphi_r$ and used $\min \frac{1}{n}\sum_i \theta_i - \frac{1}{n}\sum_r \varphi_r$ as the objective function. At the optimal solution of (42), one can multiplicatively decompose the overall Pareto-Koopmans efficiency as $\tau_{\mathrm{PK}}(x^0, y^0) = \mathrm{PK}_x(x^0, y^0).\ \mathrm{PK}_y(x^0, y^0)$, where $\mathrm{PK}_x(x^0, y^0) = \frac{1}{n}\sum_i \theta_i^*$ is the input efficiency and $\mathrm{PK}_y(x^0, y^0) = 1/\frac{1}{m}\sum_r \varphi_r^*$ is the output efficiency of the firm.

9 Conclusion

Over the decades since it was introduced in the Operations Research literature, new DEA models have been formulated to measure cost and profit efficiency as well as to test various characteristics of the technology like productivity change over time, capacity utilization, benefits of merger or break up of firms among many other areas of application in economics. In fact, DEA can now legitimately be described as a full-blown nonparametric approach to production analysis and goes way beyond merely evaluating technical efficiency.[21] Given its scope, this chapter deals only with measurement of different kinds of technical efficiency. Finally, as noted at the

[20]A somewhat different measure of Pareto-Koopmans efficiency is the Range Adjusted measure (RAM) introduced by Cooper et al. (1999).

[21]For more detailed exposition of DEA as a nonparametric approach to neoclassical production economics see Färe et al. (1994) and Ray (2004). Cooper et al. (2007) deals with most of the topics from an OR perspective.

beginning one can generate empirical distributions of the frontier output at each input bundle through bootstrapping to create upper and lower bounds on the measured technical efficiency of each firm in the sample.[22]

References

Afriat, S. 1972. Efficiency estimation of production functions. *International Economic Review* 13 (3): 568–598.

Aigner, D.J., and S.F. Chu. 1968. On estimating the industry production function. *American Economic Review* 58 (4): 826–839.

Aigner, D.J., C.A.K. Lovell, and P. Schmidt. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6 (1): 21–37.

Aparicio, J., J.T. Pastor, and S.C. Ray. 2013. An overall measure of technical inefficiency at the firm and at the industry level: The 'lost profit on outlay'. *European Journal of Operational Research* 226 (1): 154–162.

Banker, R.D. 1984. Estimating the most productive scale size using data envelopment analysis. *European Journal of Operational Research* 17 (1): 35–44.

Banker, R.D. 1993. Maximum likelihood, consistency, and data envelopment analysis: A statistical foundation. *Management Science* 39 (10): 1265–1273.

Banker, R.D., and A. Maindiratta. 1988. Nonparametric analysis of technical and allocative efficiencies in production. *Econometrica* 56 (5): 1315–1332.

Banker, R.D., and R.M. Thrall. 1992. Estimating most productive scale size using data envelopment analysis. *European Journal of Operational Research* 62: 74–84.

Banker, R.D., A. Charnes, and W.W. Cooper. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science* 30 (9) (September): 1078–1092.

Chambers, R.G., Y. Chung, and R. Färe. 1996. Benefit and distance functions. *Journal of Economic Theory* 70 (2): 407–419.

Chambers, R.G., Y. Chung, and R. Färe. 1998. Profit, directional distance functions, and Nerlovian efficiency. *Journal of Optimization Theory and Applications* 98: 351–364.

Charnes, A., and W.W. Cooper. 1968. Programming with Linear fractional functionals. *Naval Research Logistics Quarterly* 15: 517–522.

Charnes, A., W.W. Cooper, and E. Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2 (6): 429–444.

Charnes, A., W.W. Cooper, and E. Rhodes. 1979. Short communication: Measuring the efficiency of decision making units. *European Journal of Operational Research* 3 (4): 339.

[22]For details on how to bootstrap for DEA see Simar and Wilson (1998).

Cooper, W.W., L. Seiford, and K. Tone. 2007. *Data envelopment analysis: A comprehensive text with models, applications, references and DEA-solver software*, 2nd ed. Norwell, MA: Kluwer Academic Publishers.

Cooper, W.W., R.G. Thompson, and R.M. Thrall. 1996. Introduction: Extensions and new developments in DEA. *Annals of Operations Research* 66: 3–45.

Cooper, W.W., S.K. Park, K.S., and J.T. Pastor. 1999. RAM: A range adjusted measure of inefficiency for use with additive models, and relations to other models and measures in DEA. *Journal of Productivity Analysis* 11 (1): 5–42.

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

Deprins, D., L. Simar, and H. Tulkens. 1984. Labor-efficiency in post offices. In *The performance of public enterprises: Concepts and measurement*, ed. M. Marchand, P. Pestieau, and H. Tulkens, 243–267. North Holland: Elsevier Science Publications B. V.

Diewert, E., and C. Parkan. 1983. Linear programming tests of regularity conditions for production functions. In *Quantitative studies in production and prices*, ed. W. Eichorn, R. Henn, K. Neumann, and R.W. Shephard. Würzburg: Physica-Verlag.

Färe, R., and C.A.K. Lovell. 1978. Measuring the technical efficiency of production. *Journal of Economic Theory* 19 (1): 150–162.

Färe, R., S. Grosskopf, and C.A.K. Lovell. 1994. *Production frontiers.* Cambridge: Cambridge University Press.

Färe, R., S. Grosskopf, and C.A.K. Lovell. 1985. *The measurement of efficiency of production.* Boston: Kluwer-Nijhoff.

Farrell, M.J. 1957. The measurement of technical efficiency. *Journal of the Royal Statistical Society Series A (General)* 120 (3): 253–281.

Farrell, M.J., and M. Fieldhouse. 1962. Estimating efficient production functions under increasing returns to scale. *Journal of the Royal Statistical Society Series A (General)* 125 (2): 252–267.

Ferguson, C.E. 1969. *Neoclassical theories of production and distribution.* Cambridge: Cambridge University Press.

Førsund, F.R. 1996. On the calculation of scale elasticity in DEA Models. *Journal of Productivity Analysis* 7: 213–224.

Førsund, F.R., and L. Hjalmarsson. 2004. Calculating scale elasticity in DEA models. *Journal of the Operational Research Society* 55 (10): 1023–1038.

Frisch, R. 1965. *Theory of production*. Chicago: Rand McNally and Company.

Greene, W. 1980. Maximum likelihood estimation of econometric frontier functions. *Journal of Econometrics* 13: 27–56.

Hanoch, G. 1970. Homotheticity in joint production. *Journal of Economic Theory* 2: 423–426.

Hanoch, G., and M. Rothschild. 1972. Testing the assumption of production theory: A nonparametric approach. *Journal of Political Economy* 80 (2): 256–275.

Kerstens, K., and P. Vanden Eeckaut. 1999. Estimating returns to scale using non-parametric deterministic technologies: A new method based on goodness-of-fit. *European Journal of Operational Research* 113 (1): 206–214.

Koopmans, T.J. 1951. Analysis of production as an efficient combintion of activities. In *Activity analysis of production and allocation*, ed. T.J. Koopmans. Cowles Foundation.

Kumbhakar, S., and C.A.K. Lovell. 2000. *Stochastic frontier analysis*. New York: Cambridge University Press.

Luenberger, D.G. 1992. Benefit functions and duality. *Journal of Mathematical Economics* 21: 115–145.

McFadden, D. 1978. Cost, revenue, and profit functions. In *Production economics: A dual approach to theory and applications volume I: The theory of production*, ed. M. Fuss and D. McFadden, 3–109. New York: North-Holland.

Panzar, J.C., and R.D. Willig. 1977. Economies of scale in multi-output production. *Quarterly Journal of Economics* XLI: 481–493.

Pastor, J.T., J.L. Ruiz, and I. Sirvent. 1999. An enhanced DEA Russell-graph efficiency measure. *European Journal of Operational Research* 115: 596–607.

Podinovski, V. 2004a. Local and global returns to scale in performance measurement. *Journal of the Operational Research Society* 55: 170–178.

Podinovski, V. 2004b. Efficiency and global scale characteristics on the "no free lunch" assumption only. *Journal of Productivity Analysis* 22: 227–257.

Podinovski, V. 2017. Returns to scale in convex production technologies. *European Journal of Operational Research* 258: 970–982.

Portela, M.C.A.S., and E. Thanassoulis. 2005. Profitability of a sample of Portuguese bank branches and its decomposition into technical and allocative components. *European Journal of Operational Research* 162 (3): 850–866.

Ray, S.C. 2004. *Data envelopment analysis: Theory and techniques for economics and operations research*. New York: Cambridge University Press.

Ray, S.C. 2007. Shadow profit maximization and a measure of overall inefficiency. *Journal of Productivity Analysis* 27: 231–236.

Ray, S.C. 2010. A one-step procedure for returns to scale classification of decision making units in data envelopment analysis. Working Paper 2010-07, University of Connecticut Economics.

Ray, S.C. 2019. The Transformation function, technical efficiency, and the CCR Ratio. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2019.02.014.

Ray, S.C., and A. Ghose. 2014. Production efficiency in Indian agriculture: An assessment of the post green revolution years. *Omega* 44 (2014): 58–69.

Ray, S.C., and Y. Jeon. 2009. Reputation and efficiency: A non-parametric assessment of America's top-rated MBA programs. *European Journal of Operational Research* 189 (2008): 245–268.

Richmond, J. 1974. Estimating the efficiency of production. *International Economic Review* 15: 515–521.

Schmidt, P. 1976. On the statistical estimation of parametric frontier production functions. *Review of Economics and Statistics* 58: 238–239.

Seiford, L., and J. Zhu. 1999. An investigation of returns to scale in data envelopment analysis. *Omega, International Journal of Management Science* 27: 1–11.

Shephard, R.W. 1953. *Cost and production functions.* Princeton: Princeton University Press.

Shephard, R.W. 1970. *Theory of cost and production functions.* Princeton: Princeton University Press.

Simar, L., and P. Wilson. 1998. Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science* 44 (11): 49–61.

Starrett, D.A. 1977. Measuring returns to scale in the aggregate, and the scale effect of public goods. *Econometrica* 45 (6): 1439–1455.

Tone, K. 2001. A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research* 130: 498–509.

Varian, H.R. 1984. The nonparametric approach to production analysis. *Econometrica* 52 (3): 579–597.

Zhu, J. 2003. *Quantitative models for performance evaluation and benchmarking: Data envelopment analysis with spreadsheets and DEA excel solver*. Boston: Kluwer Academic Press.

Ranking Methods Within Data Envelopment Analysis

Nicole Adler and Nicola Volta

1 Introduction

Decision-making units (DMUs) within the data envelopment analysis (DEA) context are assessed based on multiple inputs and outputs, under non-parametric assumptions, which means that the production function remains unknown. A linear program is solved per DMU and the weights assigned to each linear aggregation are individual to the DMU. The weights are chosen so as to show the specific DMU in as positive a light as possible, under the restriction that no other DMU, given the same weights, is more than 100% efficient. Consequently, a Pareto frontier is delineated by specific DMUs on the boundary envelope of input-output variable space. The frontier is considered a sign of relative efficiency, which has been achieved by at least one DMU by definition. The choice of weights is highly flexible which, in general, is considered an advantage of the approach.

Many theoretical papers in the field have adapted the original set of DEA models (Charnes et al. 1978, 1985b; Banker et al. 1984) in order to handle questions that have arisen in practice. One adaptation has

N. Adler (✉)
Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: msnic@huji.ac.il

N. Volta
Cranfield University, Bedford, UK
e-mail: n.volta@cranfield.ac.uk

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_7

been in the field of ranking DMUs. The basic DEA results categorize the DMUs into two sets; those that are efficient and define the Pareto frontier and those that are inefficient. In order to rank all the DMUs, another approach or modification is required. Often decision-makers are interested in a complete ranking, beyond the dichotomized classification, in order to refine the evaluation of the units. Moreover, one problem that has been discussed frequently in the literature has been the lack of discrimination in DEA applications, in particular when there are insufficient DMUs or the number of inputs and outputs is relatively high in comparison to the number of units. This is an additional reason for the growing interest in complete ranking techniques. Furthermore, fully ranking units is an established approach in the social sciences, in general (see Young and Hamer 1987) and in multiple-criteria decision-making in particular. It should be noted that the methods discussed here could be considered post-analyses since they do not replace the standard DEA models rather provide added value. However, we also note that complete ranking is only relevant for specific contexts and not in all cases.

This chapter describes the ranking methods developed in the literature and since many articles have been published in this field, we have grouped them into seven basic areas. The methods are classified by several criteria and are not mutually exclusive. After specifying the DEA method mathematically in Sect. 2, we discuss the super-efficiency technique in Sect. 3, first published in Andersen and Petersen's paper of 1993, in which DMUs are ranked through the exclusion of the unit being scored from the linear program. In Sect. 4, we discuss the evaluation of DMUs through benchmarking, an approach originating in Torgersen et al. (1996), in which an efficient unit is highly ranked if it appears frequently in the reference sets of inefficient DMUs. In Sect. 5, we present the cross-efficiency technique, which was first suggested by Sexton et al. (1986), whereby the DMUs are both self and peer evaluated. In Sect. 6, we present the potential for ranking DMUs using a common weights approach, first discussed in Roll et al. (1991). In Sect. 7, we analyze the application of multivariate statistical tools in combination with DEA, including discriminant analysis and principal component analysis. In Sect. 8, we discuss papers that cross multi-criteria decision-making methods with the underlying concepts of DEA. Section 9 presents the ranking of inefficient DMUs, which in general is not considered in the previous sections. Finally, Sect. 10 presents the results of the various methodologies on a dataset of Higher Education Institutions (HEI) located

in the UK, and Sect. 11 draws conclusions and a summary of the various approaches to ranking. We note that this review draws from Adler et al. (2002) with extensions and additions where deemed relevant.

2 The Data Envelopment Analysis Model

DEA is a mathematical model that estimates the relative efficiency of DMU with multiple inputs and outputs but with no obvious production function in order to aggregate the data in its entirety. Relative efficiency is defined as the ratio of total weighted output to total weighted input. By comparing n units with s outputs denoted by $y_{rk}, r = 1, \ldots, s$ and m inputs denoted by $x_{ik}, i = 1, \ldots, m$, the efficiency measure for DMU k is:

$$h_k = \underset{u_r, v_i}{\text{Max}} \frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}}$$

where the weights, u_r and v_i, are non-negative. A second constraint requires that the same weights, when applied to all DMUs, do not provide any unit with an efficiency estimate greater than one. This condition appears in the following set of constraints:

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \quad \text{for } j = 1, \ldots, n$$

The efficiency ratio ranges from zero to one, with DMU k being considered relatively efficient if receiving a score of one. Thus, each unit will choose weights so as to maximize self-efficiency, given the constraints.

The result of the DEA is the determination of the hyperplanes that define an envelope surface or Pareto frontier. DMUs that lie on the surface determine the envelope and are deemed efficient, while those that do not are deemed inefficient. The formulation described above can be translated into a linear program, which can be solved relatively easily and a complete DEA solves n linear programs, one per DMU.

$$
\begin{aligned}
h_k &= \text{Max} \sum_{r=1}^{s} u_r y_{rk} \\
\text{s.t.} \sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} u_r y_{rj} &\geq 0 \qquad \text{for } j = 1, \ldots, n \\
\sum_{i=1}^{m} v_i x_{ik} &= 1 \\
u_r &\geq 0 \qquad \text{for } r = 1, \ldots, s \\
v_i &\geq 0 \qquad \text{for } i = 1, \ldots, m
\end{aligned}
\tag{1}
$$

Model (1), often referred to as the CCR model (Charnes et al. 1978), assumes that the production function exhibits constant returns to scale. The BCC (Banker et al. 1984) model adds an additional constant variable, c_k, in order to permit variable returns to scale, as shown in Model (2).

$$
\begin{aligned}
h_k &= \text{Max} \sum_{r=1}^{s} u_r y_{rk} + c_k \\
\text{s.t.} \sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} u_r y_{rj} - c_k &\geq 0 \qquad \text{for } j = 1, \ldots, n \\
\sum_{i=1}^{m} v_i x_{ik} &= 1 \\
u_r &\geq 0 \qquad \text{for } r = 1, \ldots, s \\
v_i &\geq 0 \qquad \text{for } i = 1, \ldots, m
\end{aligned}
\tag{2}
$$

It should be noted that units defined as efficient in the CCR input-minimization are the same as those defining the Pareto frontier of the output-maximized formulations, which is not necessarily the case for the results of the BCC model.

2.1 The Dual Program of the CCR Model

If a DMU proves to be inefficient, a combination of other, efficient units can produce either greater output for the same composite of inputs, use fewer inputs to produce the same composite of outputs or some combination of the two. A hypothetical decision-making unit, k', can be composed as an aggregate of the efficient units, referred to as the efficient reference set for inefficient unit k. The solution to the dual problem of the linear program directly computes the multipliers required to compile k'.

$$\begin{aligned} &\text{Min} f_k \\ &\text{s.t.} \sum_{j=1}^{n} L_{kj} x_{ij} + f_k x_{ik} \geq 0 \qquad \text{for } i = 1, \ldots, m \\ &\sum_{j=1}^{n} L_{kj} y_{rj} \geq y_{rk} \qquad \text{for } r = 1, \ldots, s \\ &L_{kj} \geq 0 \qquad \text{for } j = 1, \ldots, n \end{aligned} \tag{3}$$

In the case of an efficient DMU, all dual variables will equal zero except for L_{kk} and f_k, which reflect the unit k's efficiency, both of which will equal one. If DMU k is inefficient, f_k will equal the ratio solution of the primal problem. The remaining variables, L_{kj}, if positive, represent the multiples by which unit k's inputs and outputs should be multiplied in order to compute the composite, efficient DMU k'. L_{kj} thus defines the section of the frontier to which unit k is compared.

2.2 The Slack-Adjusted CCR Model

In the slack-adjusted DEA models, see for example model (4), a weakly efficient DMU will now be evaluated as inefficient, due to the presence of input- and output-oriented slacks s_i and σ_r, respectively.

$$\begin{aligned} &\text{Min} f_k - \varepsilon \left(\sum_{i=1}^{m} s_i + \sum_{r=1}^{s} \sigma_r \right) \\ &\text{s.t.} \sum_{j=1}^{n} L_{kj} x_{ij} + f_k x_{ik} - s_i = 0 \qquad \text{for } i = 1, \ldots, m \\ &\sum_{j=1}^{n} L_{kj} y_{rj} - \sigma_r = y_{rk} \qquad \text{for } r = 1, \ldots, s \\ &L_{kj}, s_i, \sigma_r \geq 0 \qquad \text{for } j = 1, \ldots, n \end{aligned} \tag{4}$$

where ε is a positive non-Archimedean infinitesimal. In general, this formulation creates multiple solutions as a function of the computerized value of ε hence is normally solved in two stages, first to estimate the value of f_k and subsequently to estimate the slack values given f_k^*, the optimal value.

2.3 The Additive Model

An alternative formulation proposed by Charnes et al. (1985b), utilizes slacks alone in the objective function. This model is used in both the benchmarking approach and the measure of inefficiency dominance developed in Sects. 4 and 9, respectively.

$$\begin{aligned} &\text{Min} - \sum_{i=1}^{m} s_i - \sum_{r=1}^{s} \sigma_r \\ &\text{s.t.} \sum_{j=1}^{n} L_{kj} x_{ij} - s_i = -x_{ik} \qquad \text{for } i = 1, \ldots, m \\ &\qquad \sum_{j=1}^{n} L_{kj} y_{rj} - \sigma_r = y_{rk} \qquad \text{for } r = 1, \ldots, s \\ &\qquad L_{kj}, s_i, \sigma_r \geq 0 \qquad \text{for } j = 1, \ldots, n \end{aligned} \tag{5}$$

In order to avoid large variability in the weights for all DEA models, bounds have been added through assurance regions (Thompson et al. 1986, 1990, 1992), cone ratios (Charnes et al. 1990), range adjusted measures (Cooper et al. 1999) and bounded adjusted measures (Cooper et al. 2011). This, in turn, increases the differentiability among the unit scores by reducing the number of efficient DMUs. In the extreme case, the weights will be reduced to a single set of common weights and the units will be fully ranked. However, the weight constrained literature is not discussed in this chapter since the concept does not strive, nor does it generally succeed, in reaching a complete ranking of DMUs.

3 Super-Efficiency Ranking Techniques

Andersen and Petersen (1993) developed a specific model for ranking efficient units. The methodology enables an efficient unit k to achieve an efficiency score greater than one by removing the kth constraint in the primal formulation, as shown in model (6). This is a second stage analysis that is performed after the standard first stage categorizing variables into the two sets and is applied only to those DMUs deemed efficient.

$$
\begin{aligned}
h_k = \text{Max} & \sum_{r=1}^{s} u_r y_{rk} \\
\text{s.t.} & \sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} u_r y_{rj} \geq 0 \qquad \text{for } j = 1, \ldots, n, j \neq k \\
& \sum_{i=1}^{m} v_i x_{ik} = 1 \\
& u_r \geq \varepsilon \qquad \text{for } r = 1, \ldots, s \\
& v_i \geq \varepsilon \qquad \text{for } i = 1, \ldots, m
\end{aligned}
\tag{6}
$$

The dual formulation of the super-efficient model, as seen in model (7), computes the distance between the Pareto frontier, evaluated without unit k, and the unit itself, i.e., for $J = \{j = 1, \ldots, n, j \neq k\}$.

$$
\begin{aligned}
& \text{Min } f_k \\
\text{s.t.} & \sum_{j \in J} L_{kj} x_{ij} \leq f_k x_{ij} \qquad \text{for } i = 1, \ldots, m \\
& \sum_{j \in J} L_{kj} y_{rj} \geq y_{rk} \qquad \text{for } r = 1, \ldots, s \\
& L_{kj} \geq 0 \qquad j \in J
\end{aligned}
\tag{7}
$$

However, there are three problematic issues with this methodology. First, Andersen and Petersen refer to the DEA objective function value as a rank score for all units, despite the fact that each unit is evaluated according to different weights. This value in fact explains the proportion of the maximum efficiency score that each unit k attained with its chosen weights in relation to a virtual unit closest to it on the frontier. Furthermore, if we assume that the weights reflect prices, then each unit has different prices for the same set of inputs and outputs which may not be relevant if we are analyzing branches of the same organization for example.

Second, the super-efficient methodology may give "specialized" DMUs an excessively high score. To avoid this problem, Sueyoshi (1999) introduced specific bounds on the weights in a super-efficient ranking model as described in Eq. (8).

$$
v_i \geq 1/(m+s) \max_j(x_{ij}) \quad \text{and} \quad u_r \geq 1/(m+s) \max_j(y_{rj}) \tag{8}
$$

Furthermore, in order to limit the super-efficient scores to a scale with a maximum of 2, Sueyoshi developed an Adjusted Index Number formulation, as shown in Eq. (9).

$$
\text{AIN}_k = 1 + (\delta_k^* - \min_{j \in E} \delta_j^*)/(\max_{j \in E} \delta_j^* - \min_{j \in E} \delta_j^*) \tag{9}
$$

where E is the set of text of efficient units and δ_k^* is the value of the objective function of a slack-based DEA model, first introduced by Tone (2001), and the min/max formulations refer to the super-efficient slack-based model.

The third problem lies with an infeasibility issue, which if it occurs, means that the super-efficient technique cannot provide a complete ranking of all DMUs. Thrall (1996) used the model to identify extreme efficient DMUs and noted that the super-efficiency model may be infeasible. Zhu (1996a), Dula and Hickman (1997), and Seiford and Zhu (1999) prove under which conditions various super-efficient DEA models are infeasible. Xue and Harker (2002) argue that the identification of strongly super-efficient units provides information and permits a form of ranking into groups. Du et al. (2010) point out that unlike the radial super-efficiency models, additive super-efficiency models are always feasible. Additional modifications to the variable returns-to-scale DEA models have been suggested by Lovell and Rouse (2003), Chen (2004), Chen (2005), Ray (2008), Cook et al. (2009), Johnson and McGinnis (2009), Chen and Liang (2011), Lee et al (2011), Chen et al. (2013), and Guo et al. (2017).

Despite these drawbacks, possibly because of the simplicity of the concept, many published papers have used this approach. For example, Hashimoto (1997) developed a DEA super-efficient model with assurance regions in order to rank the DMUs completely. Using model (10), Hashimoto avoided the need for compiling additional preference information in order to provide a complete ranking of the n candidates.

$$\begin{aligned} h_k = \text{Max} & \sum_{r=1}^{s} u_r y_{rk} \\ \text{s.t.} & \sum_{r=1}^{s} u_r y_{rj} \leq 1 && \text{for } j = 1, \ldots, n, j \neq k \\ & u_r - u_{r+1} \geq \varepsilon && \text{for } r = 1, \ldots, s-1 \\ & u_s \geq \varepsilon \\ & u_r - 2u_{r+1} + u_{r+2} \geq 0 && \text{for } r = 1, \ldots, s-2 \end{aligned} \tag{10}$$

where u_r is the sequence of weights given to the rth place vote (whereby each voter selects and ranks the top t candidates). The use of assurance regions avoids the specialization pitfall of the standard super-efficiency model. Indeed, Banker and Chang (2006) argue that super-efficiency is useful for outlier identification rather than ranking and Banker et al. (2017) argue that issues with reliable ranking are due for the most-part to the relatively small number of DMUs.

4 Benchmark Ranking Method

Torgersen et al. (1996) achieve a complete ranking of efficient DMUs by measuring their importance as a benchmark for inefficient DMUs. The benchmarking measure is evaluated in a two-stage procedure, whereby the additive model developed in Charnes et al. (1985b) are first used to evaluate the value of the slacks. The set of efficient units, V, is identified as those units whose slack values equal zero. In the second stage, model (11) is applied to all DMUs.

$$\begin{aligned} &\tfrac{1}{E_k} = \text{Max} f_{\text{k}} \\ &\text{s.t.} \sum_{j \in V} L_{kj} x_{ij} - s_{ik} = -x_{ik} \qquad \text{for } i = 1, \ldots, m \\ &\quad \sum_{j \in V} L_{kj} y_{rj} - f_k y_{rk} - \sigma_{rk} = 0 \qquad \text{for } r = 1, \ldots, s \\ &\quad \sum_{j \in V} L_{kj} = 1 \end{aligned} \tag{11}$$

In order to rank the efficient DMUs and evaluate which are of particular importance to the industry, the benchmarking measure aggregates the individual reference weights as shown in Eq. 12.

$$\begin{aligned} &\rho_k^r \equiv \frac{\sum_{j=1}^{n} L_{jk} \left(y_{rj}^P - y_{rj} \right)}{y_r^P - y_r} \qquad \text{for } k = 1, \ldots, V,\ r = 1, \ldots, s \\ &\text{where } y_{rj}^P = \frac{y_{rj}}{E_j} + \sigma_{rj} \end{aligned} \tag{12}$$

For each efficient DMU k, the benchmark ρ_k^r measures the fraction of total aggregated potential increase in output r, over which k acts as a reference. The efficient units together define the entire potential within each output variable. An average value of ρ_k is calculated in order to rank all efficient DMUs completely. Torgersen et al. (1996) apply their technique to a set of unemployment offices in Norway and show that, in certain cases, different conclusions were reached compared to that of Andersen and Petersen's super-efficiency method (see Sect. 3), due to the outlier problem of the latter technique. Additional papers extending the benchmarking approach include Zhu (2000) and Rezaeiani and Foroughi (2018).

This is somewhat similar to the results reported in Sinuany-Stern et al. (1994), in which an efficient DMU is highly ranked if it is chosen as a

useful target by many other inefficient units. The technique developed in this paper is applied to all DMUs in two stages. In the first stage, the efficient units are ranked by simply counting the number of times they appear in the reference sets of inefficient units, an idea first developed in Charnes et al. (1985a). The inefficient units are then ranked, in the second stage, by counting the number of DMUs that need to be removed from the analysis before they are considered efficient. However, a complete ranking cannot be ensured since many DMUs may receive the same ranked score.

5 Cross-Efficiency Ranking Methods

The cross-evaluation matrix was first developed by Sexton et al. (1986), inaugurating the subject of ranking in DEA. Indeed, as Doyle and Green (1994) argue, decision-makers do not always possess a reasonable mechanism from which to choose assurance regions, thus they recommend the cross-evaluation matrix for ranking units. The cross-efficiency method estimates the efficiency score of each DMU n times, using the set of optimal weights evaluated by the n LPs. The results of all the DEA cross-efficiency scores are summarized in a cross-efficiency matrix as shown in Eq. (13).

$$h_{kj} = \frac{\sum_{r=1}^{s} u_{rk} y_{rj}}{\sum_{i=1}^{m} v_{ik} x_{ij}} \quad \text{for } k = 1, \ldots, n, \; j = 1, \ldots, n \tag{13}$$

Thus, h_{kj} represents the score given to unit j in the DEA run of unit k, i.e., unit j is evaluated by the weights of unit k. Note that all the elements in the matrix lie between zero and one, $0 \leq h_{kj} \leq 1$, and the elements in the diagonal, h_{kk}, represent the standard DEA efficiency score, whereby $h_{kk} = 1$ for efficient units and $h_{kk} < 1$ for inefficient units. Furthermore, if the weights of the LP are not unique, a goal-programming technique may be applied to choose between the optimal solutions. According to Sexton et al. (1986), the secondary goals could be, for example, either aggressive or benevolent. In the aggressive context, DMU k chooses among the optimal solutions such that it maximizes self-efficiency and at a secondary level minimizes the other DMUs cross-efficiency levels. The benevolent secondary objective would be to maximize all DMUs cross-efficiency rankings. See also Oral et al. (1991) for methods of evaluating the goal programs.

The cross-efficiency ranking method in the DEA context utilizes the results of the cross-efficiency matrix h_{kj} in order to rank-scale the units. Let us define, $\bar{h}_k = \sum_{j=1}^{n} h_{kj}/n$, as the average cross-efficiency score given to unit k. Averaging, however, is not the only possibility, as the median, minimum or variance of scores could also be applied. Green et al. (1996) provide further detailed suggestions to help avoid such problems as the lack of independence of irrelevant alternatives. It could be argued that $\bar{h}_k$, or an equivalent, is more representative than h_{kk}, the standard DEA efficiency score, since all the elements of the cross-efficiency matrix are considered, including the diagonal. Furthermore, all the units are evaluated with the same sets of weight vectors. Consequently, the $\bar{h}_k$ score better represents the unit evaluation since it measures the overall ratios over all the runs of all the units. The maximum value of $\bar{h}_k$ is 1, which occurs if unit k is efficient in all the runs, i.e., all the units evaluate unit k as efficient. In order to rank the units, we simply assign the unit with the highest score a rank of 1 and the unit with the lowest score a rank of n. While the DEA scores, h_{kk}, are non-comparable, since each uses different weights, the $\bar{h}_k$ score is comparable because it utilizes the weights of all the units equally. However, this is also the drawback of the technique, since the evaluation subsequently loses its connection to the multiplier weights.

Furthermore, Doyle and Green (1994) developed the idea of a "maverick index" within the confines of cross-efficiency. The index measures the deviation between h_{kk}, the self-appraised score, and the unit's peer scores, as shown in Eq. (14).

$$M_k = \frac{h_{kk} - e_k}{e_k} \quad \text{where } e_k = \frac{1}{(n-1)} \sum_{j \neq k} h_{kj} \tag{14}$$

The higher the value of M_k, the more the DMU could be considered a maverick. Doyle and Green (1994) argue that this score can go hand-in-hand with the benchmarking process (Sect. 4), whereby DMUs considered efficient under self-appraisal but fail to appear in the reference sets of inefficient DMUs will generally achieve a high M_k score. Those achieving a low score are generally considered all-round performers and the DMUs are frequently both self and peer efficient. Liang et al. (2008) extend Doyle and Green (1994) by proposing alternative secondary goals including minimizing the total deviation from the ideal point, the maximum deviation from the efficiency score or the mean absolute deviation.

6 Common Weights for Ranking Decision-Making Units

One method for ensuring an almost complete ranking of DMUs is the estimation of a common set of weights. This of course changes the meaning of data envelopment analysis in which each DMU is permitted to determine an individual set of weights, provided no other DMU receives a score greater than one for the chosen set of weights. However, a common set of weights is the standard approach in most engineering and economic analyses such as stochastic frontier models. The approach also enables a comparison and ranking of all DMUs, irrespective of whether they are DEA efficient or not. As suggested in Roll et al. (1991), assuming a uniformity of operational procedures may be relevant when analyzing a set of branches of a firm, and a comparison between a standard set of DEA weights and common weights could indicate special circumstances for a specific DMU.

Clearly, there are many potential methods for assessing a common set of weights. Roll et al. (1991) suggest three potential approaches, including (i) a standard DEA analysis and then the choice of average weights per factor or an alternative central measure; (ii) applying bounds on weights; and (iii) employing additional, external information on the importance of factors and setting higher or lower bounds accordingly. Using an example of highway maintenance stations in the USA, their common weight set was based on averages between bound values, which led to a complete ranking of units.

Since then, many papers have been published offering a myriad of possibilities for estimating a common set of weights. For example, Kao and Hung (2005) propose a compromise solution method based on an ideal point.

$$\begin{aligned} &\min D_P = \left[\sum_{j=1}^{n}\left(E_j^* - E_j(u,v)\right)^p\right]^{1/p} \\ &\text{s.t. } E_j(u,v) = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \qquad \text{for } j = 1,\ldots,n \\ &\quad u_r, v_i \geq \varepsilon > 0, \qquad \text{for } r = 1,\ldots,s,\ i = 1,\ldots,m \end{aligned} \tag{15}$$

where E_j^* represents the target or ideal solution and p represents the distance parameter whereby $1 \leq p < \infty$. Formulation (15) requires solving a non-linear objective function and linear constraints. Multiple potential ranks are possible as a function of the value of parameter p although Kao and Hung argue that $p=2$ is probably the most appropriate because (i) the

objective function then minimizes the path to the target solution based on Euclidean space and (ii) is unique.

A more recent paper by Kritikos (2017) develops a common set of weights model based on five dummy DMUs that span the production possibility set by creating ideal and anti-ideal virtual observations. The model draws from the technique for order preference by similarity to the ideal solution generally known as TOPSIS (Hwang and Yoon 1981). Kritikos develops an objective function which simultaneously minimizes the geometric distance from the positive ideal solution and maximizes the distance to the negative ideal solution in order to estimate a common set of weights. The common weights then create a fully ranked set of DMUs.

7 Combining Statistical Approaches with Data Envelopment Analysis

Alternative approaches to setting common weights involve the use of statistical techniques in conjunction with DEA to achieve a complete ranking. It should be noted that DEA is a methodology directed toward frontiers rather than central tendencies. Instead of trying to fit regression planes through the center of the data, DEA floats a piecewise linear surface (the efficient frontier) that rests on top of the observations. DEA optimizes and thereby focuses on each unit separately, while regression is a parametric approach that fits a single function to the data collected on the basis of average behavior that requires the functional form to be pre-specified. On the other hand, in DEA the values of the weights differ from unit to unit and while this flexibility in the choice of weights characterizes the DEA model, different weights cannot generally be used for ranking because these scores are obtained from different comparison (peer) groups for each inefficient DMU.

7.1 Discriminant Analysis of Ratios for Ranking

Sinuany-Stern and Friedman (1998) developed a technique in which discriminant analysis of ratios was applied to DEA (DR/DEA). Instead of considering a linear combination of the inputs and outputs in one equation, as in traditional discriminant analysis of two groups, a ratio function is constructed as a linear combination of the inputs and a linear combination of the outputs. In some ways, this ratio function is similar to the DEA efficiency ratio; however, while DEA provides weights for the inputs and

outputs, which vary from unit to unit, DR/DEA provides common weights for all units. In principle, DR/DEA determines the weights such that the ratio score function discriminates optimally between two groups of DMUs on a one-dimensional scale (in this case, efficient and inefficient units pre-determined by DEA). The ratio, T_j, and the arithmetic means of the ratio scores of the efficient and inefficient groups are:

$$T_j = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}}, \quad j = 1, \ldots, n$$

$$\text{where } \overline{T}_1 = \sum_{j=1}^{n_1} \frac{T_j}{n_1} \quad \text{and} \quad \overline{T}_2 = \sum_{j=n_1+1}^{n} \frac{T_j}{n_2}$$

and n_1 and n_2 are the number of efficient and inefficient units in the DEA model, respectively. The weighted mean of the entire n units ($n = n_1 + n_2$) is denoted by: $\overline{T} = \frac{n_1\overline{T}_1 + n_2\overline{T}_2}{n}$.

The problem is to find the common weights v_i and u_r such that the ratio of the between-group variance of T, ($SS_B(T)$) and the within group variance of T, ($SS_W(T)$) will be maximized, as shown in model (16).

$$\begin{aligned} &\max_{u_r, v_i} \lambda = \max_{u_r, v_i} \frac{SS_B(T)}{SS_W(T)} \\ &SS_B(T) = n_1(\overline{T}_1 - \overline{T})^2 + n_2(\overline{T}_2 - \overline{T})^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{T}_1 - \overline{T}_2)^2 \\ &SS_W(T) = \sum_{j=1}^{n_1} (T_j - \overline{T}_1)^2 + \sum_{j=n_1+1}^{n} (T_j - \overline{T}_2)^2 \end{aligned} \tag{16}$$

DR/DEA constructs the efficiency score for each unit j as T_j, the ratio between the composite output and the composite input. Hence the procedure rank scales the DMUs so that the unit with the highest score receives rank 1 and the unit with the lowest score ranks n. If any weight is negative, then non-negativity constraints ought to be added to the optimization problem. To solve this problem, it is necessary to apply a non-linear optimization algorithm; however, there is no guarantee that the solution found is globally optimal.

7.2 Principal Component Analysis for Improved Discrimination

Another attempt to improve the discriminating power of DEA is developed in Adler and Golany (2001, 2002), where principal component analysis is utilized to reduce the number of inputs/outputs, thus reducing the problem of dimensionality. Lack of discrimination, often defined as the curse of dimensionality, means that a large number of DMUs are incorrectly classified as efficient due to the overestimation bias. Principal components, a methodology that produces uncorrelated linear combinations of the original variables, ought to improve discrimination in DEA with minimal loss of information. This approach assumes that separation of variables representing similar themes, such as quality or environmental measures, and the removal of principal components with little or no explanatory power, improves the categorization of efficient and inefficient DMUs.

Let the random vector $X = [X_1, X_2, \ldots, X_p]$ possess the covariance matrix V with eigenvalues $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_p \geq 0$ and normalized eigenvectors $l_1, l_2, \ldots, l_p$. Consider the linear combinations, where the superscript t represents the transpose operator, as specified in Eq. (17). The new variables, commonly known as principal components, are weighted sums of the original data.

$$\begin{aligned} &X_{PC_i} = l_i^t X = l_{1i}X_1 + l_{2i}X_2 + \cdots + l_{pi}X_p \\ &\mathrm{Var}\left(X_{PC_i}\right) = l_i^t V l_i = \eta_i \qquad \text{for } i = 1, 2, \ldots, p \\ &\mathrm{Cov}\left(X_{PC_i}, X_{PC_k}\right) = l_i^t V l_k = 0 \qquad \text{for } i = 1, 2, \ldots, p,\ k = 1, 2, \ldots, p,\ i \neq k \end{aligned} \tag{17}$$

The principal components, $X_{PC_1}, X_{PC_2}, \ldots, X_{PC_p}$, are the uncorrelated linear combinations ranked by their variances in descending order. In order to counter bias that might occur due to differences in the magnitude of the values of the original variables, the PC transformation should be applied to the correlation matrix of the normalized variables. Principal components are computed based solely on the correlation matrix, and their development does not require a multivariate normal assumption. The complete set of principal components is as large as the original set of variables. L_x is the matrix of all l_i whose dimensions drop from $m \times m$ to $h \times m$, as PCs are dropped (X_{pc} becomes an $h \times n$ matrix). PCA-DEA is defined as a linear program in models (18).

$$\begin{aligned} &\underset{s_{pc},\sigma_{pc},\lambda}{\text{Max}}\ w^t_{Ypc}s_{pc} + w^t_{Xpc}\sigma_{pc} \\ &\text{s.t. } Y_{pc}\lambda - L_y s_{pc} = Y^a_{pc} \\ &\quad -X_{pc}\lambda - L_x\sigma_{pc} = -X^a_{pc} \\ &\quad \sigma_{pc}, s_{pc}, \lambda \geq 0 \end{aligned} \tag{18a}$$

$$\begin{aligned} &\underset{V_{pc},U_{pc}}{\text{Min}}\ V^t_{pc}X^a_{pc} - U^t_{pc}Y^a_{pc} \\ &\text{s.t. } V^t_{pc}X_{pc} - U^t_{pc}Y_{pc} \geq 0 \\ &\quad V^t_{pc}L_x \geq w^t_{Xpc} \\ &\quad U^t_{pc}L_y \geq w^t_{Ypc} \\ &\quad V_{pc} \text{ and } U_{pc} \text{ are free} \end{aligned} \tag{18b}$$

where subscript "*pc*" is the index of principal component variables; X_{pc} represents an *m by n* input matrix; Y_{pc} an *r by n* output matrix; λ a column *n*-vector of DMU weights; σ a column *m*-vector of input excess; *s* a column *r*-vector of output slack variables; w^t is a vector consisting of reciprocals of the sample standard deviations of the relevant variables. An additional constraint $e^t\lambda = 1$ can be added to (18a) corresponding to the BCC case. (18b) is the dual version of (18a). The PCA-DEA formulation is exactly equivalent to the original linear program if and only if the PCs explain 100% of the variance in the original input and output matrices. Based on Adler and Yazhemsky (2010), it would appear to be inadvisable to reduce information below 80% and this may not be sufficient to achieve a complete ranking of DMUs.

8 Multi-criteria Decision-Making

The multi-criteria decision-making (MCDM) literature was entirely separate from DEA research until 1988, when Golany combined interactive, multiple-objective linear programming and DEA. While the MCDM literature does not consider a complete ranking as their ultimate aim, they do discuss the use of preference information to further refine the discriminatory power of the models. In this manner, the decision-makers could specify which inputs and outputs should lend greater importance to the model solution. However, this could also be considered the weakness of these methods, since additional knowledge on the part of the decision-makers is required. Golany (1988), Kornbluth (1991), Thanassoulis and Dyson (1992), Golany

and Roll (1994), Zhu (1996b), and Halme et al. (1999) each incorporated preferential information into the DEA models through, for example, a selection of preferred input/output targets or hypothetical DMUs. Joro et al. (1998) and Gandibleux (2006) provide reviews of the literature on DEA and multiple-objective linear program (MOLP). A separate set of papers reflected preferential information through limitations on the values of the weights (assurance regions or cone-ratio models), which can almost guarantee a complete DMU ranking. Such papers include Thompson et al. (1986), Dyson and Thanassoulis (1988), Charnes et al. (1989, 1990), Cook and Kress (1990a, b), Thompson et al. (1990), Wong and Beasley (1990), Cook and Johnston (1992) and Green and Doyle (1995).

Cook and Kress (1990a, b, 1991, 1994) and Cook et al. (1993, 1996) argued that by imposing ratio constraints on the multipliers and replacing the infinitesimal with a lower bound thus acting as a discrimination factor, the modified DEA model can almost ensure a unique efficient DMU. For example, when considering aggregation of votes whereby y_{rj} is the number of r^{th} placed votes received by candidate j, one can define a discrimination intensity function $d(r,\varepsilon)$ and solve model (19).

$$\begin{aligned} &\text{Max } \varepsilon \\ &\text{s.t. } \sum_{r=1}^{s} u_r y_{rj} \le 1 \qquad \text{for } j = 1, \ldots, n \\ &\quad u_r - u_{r+1} - \mathrm{d}(r,\varepsilon) \ge 0 \\ &\quad u_r - \mathrm{d}(r,\varepsilon) \ge 0 \\ &\quad u_r, \varepsilon \ge 0 \end{aligned} \tag{19}$$

where $\mathrm{d}(r,\varepsilon)$ ensures that first-place votes are valued at least as highly as second-place votes and so on. The ease with which this formulation can be solved depends on the form of $\mathrm{d}(r,\varepsilon)$. Model (19) is linear if the difference between the ranks is linear, but this need not always be the case. However, as pointed out in Green et al. (1996), the form of $\mathrm{d}(r,\varepsilon)$ affects the ranking results and no longer allows DMUs to choose their weights freely. Furthermore, which mathematical function is appropriate is unclear yet important to the analysis.

Some have gone as far as to argue that DEA should be considered another methodology within MCDM, for example Troutt (1995), Li and Reeves (1999), and Sinuany-Stern et al. (2000). Troutt (1995) developed a maximin efficiency ratio model in which a set of common weights is evaluated to distinguish between efficient DMUs as shown in model (20).

$$\begin{aligned}
&\underset{u_r,v_i}{\text{Maximize}} \left(\underset{k}{\text{Minimize}} \frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}} \right) \\
&\text{s.t.} \frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}} \leq 1 \qquad \text{for all } k \\
&\sum_{r=1}^{s} u_r = 1 \\
&u_r, v_i \geq 0 \qquad \text{for all } r, i
\end{aligned} \tag{20}$$

Li and Reeves (1999) suggest utilizing multiple objectives, such as minimax and minisum efficiency in addition to the standard DEA objective function in order to increase discrimination between DMUs, as shown in the MOLP of model (21).

$$\begin{aligned}
&\text{Min } d_k \\
&\text{Min } M \\
&\text{Min} \sum_{j=1}^{n} d_j \\
&\text{s.t.} \sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} u_r y_{rj} + d_j = 0 \qquad \text{for } j = 1, \ldots, n \\
&\sum_{i=1}^{m} v_i x_{ik} = 1 \\
&M - d_j \geq 0 \\
&u_r, v_i, d_j \geq 0 \qquad \text{for all } r, i, j
\end{aligned} \tag{21}$$

The first objective is equivalent to the standard CCR model. The second objective function requires the MOLP to minimize the maximum deviation (slack variable) and the third objective is to minimize the sum of deviations. The aim is to increase discrimination, which the second and third objectives provide without promising complete ranking, similar to the assurance regions and cone-ratio approaches. On the other hand, this approach does not require additional preferential information as do other approaches. Extensions to this approach can be found in Ghasemi et al. (2014) and de Carvalho Chaves et al. (2016). Rubem et al. (2017) argue that the Li and Reeves model only finds non-dominated solutions rather than a full ranking and propose a weighted goal program combined with DEA to achieve this purpose. Model (22) translates the three goals (g_1, g_2, g_3) defined in model (21) into deviational variables and creates a complete rank although the

weights (w_1, w_2, w_3) in the objective function would still require additional information beyond that of the standard DEA modeling approach.

$$\begin{aligned}
&\text{Min } a = \left\{w_1 d_1^+ + w_3 d_3^+ + w_3 d_3^+\right\} \\
&\text{s.t. } \sum_{i=1}^{m} v_i x_{ij} = 1 \\
&\quad \sum_{r=1}^{s} u_r y_{rj} + \sum_{i=1}^{m} v_i x_{ij} + d_j = 0, \quad \text{for } j = 1, \ldots, n \\
&\quad M - d_j \geq 0, \text{for all } j \\
&\quad d_0 + d_1^- - d_1^+ \leq g_1 \\
&\quad M + d_2^- - d_2^+ \leq g_2, \text{for all } k \\
&\quad \sum_{k=1}^{n} d_k + d_3^- - d_3^+ \leq g_3 \\
&\quad u_r, v_i, d_j, d_1^-, d_1^+, d_2^-, d_2^+, d_3^-, d_3^+ \geq 0, \quad \text{for all } r, i, j
\end{aligned} \tag{22}$$

where $d_1^-, d_1^+, d_2^-, d_2^+, d_3^-, d_3^+$. represent deviational variables and M a large number.

However, it should be noted that certain researchers have argued that MCDM and DEA are two entirely separate approaches, which do not overlap. MCDM is generally applied to ex-ante problem areas where data is not readily available, especially if referring to a discussion of future technologies, which have yet to be developed. DEA, on the other hand, provides an ex-post analysis of the past from which to learn. A discussion of this topic can be found in Belton and Stewart (1999).

9 Ranking Inefficient Decision-Making Units

The majority of techniques so far discussed have not attempted to rank the inefficient DMU beyond the efficiency scores attained from the standard DEA models. However, it could be argued that comparing the scores of two inefficient units is only possible if their benchmarks draw from the same sub-set of efficient units. The super-efficiency method ranks inefficient units in the same manner as the standard DEA model. The benchmarking concept only attempts to rank DMUs identified in the standard DEA models as efficient. It should be noted that both the cross-efficiency method and the various statistical techniques do attempt to address this problem.

One concept, derived in Bardhan et al. (1996), ranks inefficient units using a measure of inefficiency dominance (MID). The measure is based on

slack-adjusted DEA models from which an overall measure of inefficiency can be computed as shown in Eq. (23).

$$0 \leq 1 - \frac{\sum_{i=1}^{m} \frac{s_r^*}{x_{ik}} + \sum_{r=1}^{s} \frac{\sigma_i^*}{y_{rk}}}{m+s} \leq 1 \quad \text{for } k = 1, \ldots, n \tag{23}$$

The MID index ranks the inefficient DMUs according to their average proportional inefficiency in all inputs and outputs. However, just as the benchmarking approach (see Sect. 4) only ranks the efficient units, the MID index only ranks the inefficient units.

10 Illustration of Ranking

For the benefit of the reader, we analyze a simple illustration comparing Higher Education Institutes (HEI) in the UK using the techniques presented in this chapter. Rankings and league tables in higher education have experienced increasing popularity over the last decade. The ranks have become commercialized and affect managerial decisions at HEI globally (Hazelkorn 2015; Ruiz and Sirvent 2016). However, despite their popularity, it has been widely acknowledged in the related literature that university rankings are controversial (De Witte and Hudrlikova 2013).

The data for our illustration has been collected from the UK's Higher Education Statistics Agency (HESA) and refers to the academic year 2015–2016. Given the scope of this section, our observations are restricted to the English universities belonging to the Russell Group. The Russell Group, founded in 1994, represents 24 of the leading universities in the UK. The Russell Group includes heterogeneous providers in terms of the number of students and courses provided. For example, it includes highly specialized providers such as the London School of Economics and Political Science (social sciences), as well as comprehensive universities such as University College London. Consequently, the rankings are only intended as a numerical exercise to highlight differences in the modeling approaches.

For the purposes of the illustration, we assume a production process in which academic and non-academic staff represent inputs while teaching and research income represent outputs. The set of HEI providers and the inputs and outputs are presented in Table 1.[1]

[1]We included in our sample 20 of the 24 members of the Russell Group. Data for Cardiff University, University of Edinburgh, University of Glasgow and Queen's University Belfast was not available.

Table 1 List of higher education providers

HEI provider	Academic staff	Non-academic staff	Teaching income (£000's)	Research income (£000's)
Imperial College of Science, Technology and Medicine	4370	3870	304,207	444,816
King's College London	4710	2880	321,926	258,526
London School of Economics and Political Science	1655	1900	183,901	50,096
Newcastle University	2750	3200	259,506	144,034
Queen Mary University of London	2185	2060	221,200	133,617
The University of Birmingham	3635	3795	337,157	175,084
The University of Bristol	3055	3185	260,885	195,478
The University of Cambridge	5825	5220	335,834	590,439
The University of Exeter	1870	2335	226,881	84,782
The University of Leeds	3345	4195	331,852	171,845
The University of Liverpool	2825	3010	282,716	121,918
The University of Manchester	5075	5495	484,856	342,200
The University of Oxford	6945	6095	346,926	676,474
The University of Sheffield	3255	3980	332,223	211,251
The University of Southampton	2995	3365	273,877	158,022
The University of Warwick	2740	3570	298,265	151,954
The University of York	1710	2190	173,032	87,320
University College London	7220	4995	483,880	661,934
University of Durham	1690	2815	183,343	76,311
University of Nottingham	3410	4115	342,568	173,529

Table 2 Radial and additive, constant and variable returns-to-scale estimates

HEI provider	RAD CCR	RAD BCC	ADD CCR	ADD BCC
Imperial College of Science, Technology and Medicine	1	1	0	0
King's College London	1	1	0	0
London School of Economics and Political Science	1.03	1	41,602	0
Newcastle University	1.14	1.13	61,579	61,167
Queen Mary University of London	1	1	0	0
The University of Birmingham	1.16	1.05	10,2554	90,387
The University of Bristol	1.12	1.11	62,851	62,609
The University of Cambridge	1	1	72,139	0
The University of Exeter	1	1	0	0
The University of Leeds	1.11	1.04	60,573	55,873
The University of Liverpool	1.09	1.02	68,129	61,559
The University of Manchester	1.03	1	30,754	0
The University of Oxford	1.04	1	163,351	0
The University of Sheffield	1.01	1	4762	0
The University of Southampton	1.16	1.14	7,5459	74,529
The University of Warwick	1.02	1	9572	0
The University of York	1.10	1	27,799	0
University College London	1	1	0	0
University of Durham	1.09	1	23,997	0
University of Nottingham	1.11	1.03	58,747	52,223

The results of the output-oriented radial (RAD) and additive (ADD) models assuming constant (CCR) and variable (BCC) returns to scale are presented in Table 2. We note that a score of one in the radial models implies efficiency and a score greater than one indicates the potential increase in income achievable were the HEI to lie on the Pareto frontier. In the additive model, a score of zero identifies efficient units and a positive value in the objective function suggests by how much the sum of incomes ought to increase in order for the HEI to be deemed efficient. The CCR models identify six efficient units whereas the BCC models identify the same six units and six additional efficient universities. Four of the six CCR radial efficient DMUs are located in London. Perhaps interestingly, the additive model results are the same except for Cambridge University, which is only weakly CCR-efficient hence should be able to increase output in the range of £72 million in order to lie on the strongly efficient frontier.

Table 3 presents the results for the aggressive and benevolent cross-efficiency specifications, the CCR and BCC super-efficiency models and the Common Set of Weights (CSW) radial, variable returns-to-scale model. In the two cross-efficiency specifications, none of the HEI providers obtain

Table 3 Cross-efficiency and super-efficiency estimates

HEI provider	Benevolent cross-efficiency	Aggressive cross-efficiency	Super-efficiency CCR	Super-efficiency BCC	CSW
Imperial College of Science, Technology and Medicine	1.16	2.79	0.93	0.90	1.26
King's College London	1.58	2.87	0.93	0.92	0.96
London School of Economics and Political Science	2.73	3.38	1.03	Infinity	0.86
Newcastle University	1.91	2.88	1.14	1.13	0.92
Queen Mary University of London	1.51	2.37	0.93	0.63	1.10
The University of Birmingham	1.94	2.86	1.16	1.05	0.91
The University of Bristol	1.61	2.75	1.12	1.11	0.99
The University of Cambridge	1.27	3.32	1.00	0.98	1.17
The University of Exeter	2.05	2.75	0.91	0.90	0.99
The University of Leeds	1.98	2.92	1.11	1.04	0.91
The University of Liverpool	2.08	2.88	1.09	1.02	0.91
The University of Manchester	1.53	2.54	1.03	0.83	1.06
The University of Oxford	1.37	3.77	1.04	0.96	1.10
The University of Sheffield	1.63	2.59	1.01	0.99	1.04
The University of Southampton	1.88	2.88	1.16	1.14	0.92
The University of Warwick	1.86	2.73	1.02	1.00	0.98
The University of York	2.01	2.93	1.10	0.74	0.92
University College London	1.15	2.76	0.85	0.75	1.24
University of Durham	2.49	3.47	1.09	0.97	0.82
University of Nottingham	1.95	2.85	1.11	1.03	0.93

a final score of one. Moreover, comparing the two sets of results, the differences are somewhat stark with Imperial College obtaining the highest rank in the benevolent specification and University College London in the aggressive model. Super-efficiency scores for inefficient units match the results presented in Table 2 (radial CCR and BCC) for the inefficient units since the model only ranks the efficient DMUs. University College London is ranked as the most efficient in the CCR case and Queen Mary University of London in the BCC case. However, we note that the BCC super-efficiency case results in one infeasible solution, namely the London School of Economics. The London School of Economics could be deemed a specialist as it consists of 27 departments and institutes alone. It could be argued that the School is an outlier, which matches the argument of Banker et al. (2017) or equally, a strongly efficient unit as argued in Xue and Harker (2002). Using the average of weights in order to estimate a common set of weights (Roll et al. 1991) leads to weights of two thirds with respect to academic staff and one third with respect to support staff. Teaching revenues receive a weight of 0.6 and research funds equal to 0.4. The results of CSW give a high score to Imperial College and University College London and low scores to the University of Durham and the London School of Economics. In other words, the application of a single fixed set of weights, for example to generate the Human Development Index (Anand and Sen 1994), creates a starkly different ranking as a result of the assumption of a single production function.

Results for the benchmarking approach (based on the radial BCC model), the measure of inefficiency dominance (MID), and principal component analysis (PCA-DEA) are presented in Table 4. The benchmarking model, which only ranks the efficient units, indicates that the universities of Exeter and Manchester are ranked first and second, respectively. This result is somewhat different to the super-efficient model because super-efficiency tends to rank the specialized DMUs highly whereas the benchmarking approach is more likely to highlight DMUs that are reasonably good with respect to all variables. Clearly, the choice of method impacts the results hence the reason for benchmarking should guide the user in choosing the most appropriate approach.

The MID model, which ranks only inefficient units, indicates that the University of Birmingham is the lowest ranked unit. Since MID accounts for the average proportional inefficiency, the results do not necessarily match the ranking of the additive BCC model precisely. The principal component analysis approach reduces both inputs and outputs to single weighted variables representing 93 and 84%, respectively, of the underlying correlation

Table 4 Benchmarking ranks, MID ranks and PCA-DEA scores

HEI provider	Benchmarking	MID	PCA-DEA CCR	PCA-DEA BCC
Imperial College of Science, Technology and Medicine	5		1.08	1.07
King's College London	7		1.11	1.10
London School of Economics and Political Science	7		1.16	1.00
Newcastle University		3	1.24	1.23
Queen Mary University of London	3		1.00	1.00
The University of Birmingham		6	1.19	1.18
The University of Bristol		2	1.20	1.19
The University of Cambridge	7		1.22	1.20
The University of Exeter	1		1.07	1.07
The University of Leeds		3	1.25	1.24
The University of Liverpool		5	1.16	1.15
The University of Manchester	2		1.11	1.10
The University of Oxford	7		1.33	1.20
The University of Sheffield	6		1.15	1.14
The University of Southampton		4	1.24	1.23
The University of Warwick	4		1.17	1.16
The University of York	7		1.25	1.19
University College London	7		1.01	1.00
University of Durham	7		1.44	1.43
University of Nottingham		1	1.21	1.20

between the two variable sets. The PCA-DEA results suggest that Queen Mary University is the highest ranked, closely followed by University College London based on the CCR approach and joined by the London School of Economics under the BCC approach. The lowest ranked universities include Oxford, Birmingham and Southampton joined by Durham in the BCC PCA-DEA model, all located outside of the London conurbation. Perhaps surprisingly, the two mostly highly considered universities in the UK, namely Oxford and Cambridge, are ranked efficient but lowest in the benchmarking approach and inefficient according to PCA-DEA due to the use of excessive staff levels in comparison with research and teaching income.

Based on the scores presented in Tables 2–4, we created ranks which are summarized in Table 5. Queen Mary, University College London

Table 5 Complete HEI ranking

HEI provider	Radial		Additive		Super-eff		Cross-eff		CSW	Bench	MID	PCA-DEA		QS ranking	THE ranking
	CCR	BCC	CCR	BCC	CCR	BCC	Benevolent	Aggressive			CCR	CCR	BCC		
Queen Mary University of London	1	1	1	1	3	1	5	1	5	3		1	2	16	12
University College London	1	1	1	1	1	3	1	7	2	7		1	2	3	4
Imperial College of Science, Technology and Medicine	1	1	1	1	4	6	2	8	1	5		1	3	4	3
The University of Exeter	1	1	1	1	2	5	17	5	8	1		1	4	19	14
The University of Warwick	3	1	3	1	8	12	10	4	10	4		3	9	9	9
The University of Sheffield	2	1	2	1	7	11	9	3	7	6		2	7	13	11
The University of Manchester	5	1	6	1	10	4	6	2	6	2		6	5	6	7
King's College London	1	1	1	1	5	7	7	11	11	7		1	6	5	6
The University of Cambridge	1	1	13	1	6	10	3	17	3	7		13	15	1	2
London School of Economics and Political Science	4	1	7	1	9	20	20	18	19	7		7	2	7	5
The University of Oxford	6	1	16	1	11	8	4	20	4	7		16	14	2	1
The University of York	9	1	5	1	14	2	16	16	15	7		5	12	17	15
University of Durham	7	1	4	1	12	9	19	19	20	7		4	19	10	10
University of Nottingham	10	3	8	2	15	14	14	9	12		1	8	13	11	18
The University of Bristol	12	6	11	6	17	17	8	6	9		2	11	11	8	8
The University of Liverpool	8	2	12	5	13	13	18	12	18		5	12	8	18	19
Newcastle University	13	7	10	4	18	18	12	13	13		3	10	16	20	20
The University of Leeds	11	4	9	3	16	15	15	15	17		3	9	18	15	17
The University of Birmingham	15	5	15	8	20	16	13	10	16		6	15	10	12	16
The University of Southampton	14	8	14	7	19	19	11	14	14		4	14	17	14	13

and Imperial receive the highest ranks across most models in general. At the opposite end of the spectrum, the universities of Southampton, Birmingham, Leeds, Durham and Newcastle consistently appear the weakest in terms of maximizing their teaching and research income given their academic and support staff resources. It would appear that most methodologies arrive at similar conclusions with respect to the highest and lowest performing HEIs but the remaining ranks are very much dependent on the underlying rationale of the individual models. Furthermore, three universities (London School of Economics, Oxford and Cambridge) show the largest changes in ranks across the models, suggesting that additional variables may be missing from the current dataset.

The last two columns of Table 5 report the relative rankings according to the 2017 QS world university ranking and the 2017 Times Higher Education ranking (THE). We note that the two indices follow additional criteria over and above the inputs and outputs used in the analysis presented here.[2] While the rankings appear similar for Imperial College and University College London, Queen Mary receives a much lower score in the QS and THE ranks compared to the results of the DEA ranking. It would appear that the commercial rankings place greater emphasis on reputation, and perhaps quality, than the efficient use of resources.

In order to visualize the universities in the dataset, Fig. 1 is created from a Co-plot model (Raveh 2000; Adler and Raveh 2008). Co-plot locates each DMU in a two-dimensional space in which the location of each DMU is determined by all variables simultaneously. In a subsequent stage, all ratios of output to input are plotted in the form of arrows and superimposed sequentially. The DMUs are exhibited as n points and the ratios are exhibited as k arrows relative to the same axis and origin. A measure of goodness-of-fit is computed and associated to each criterion separately. Co-plot is based on the integration of mapping concepts using a variant of regression analysis. Starting with a data matrix X_{nk} of n rows and k columns, Co-plot is composed of four stages. In stage 1, X_{nk} is normalized and the elements are deviations from column means divided by standard deviations. In stage 2, a measure of dissimilarity $D_{ij} \geq 0$ is computed based on the sum of absolute deviations between each pair of observations. In stage 3, the n DMUs are

[2]The Times Higher Education ranking aggregates 13 weighted performance indicators in the categories of teaching (the learning environment), research (volume, income and reputation), citations (research influence), international outlook (staff, students and research) and industry income (knowledge transfer). Similarly, the QS world university ranking is based on six weighted metrics: academic reputation, employer reputation, faculty/student ratio, citations per faculty, international faculty ratio and international student ratio.

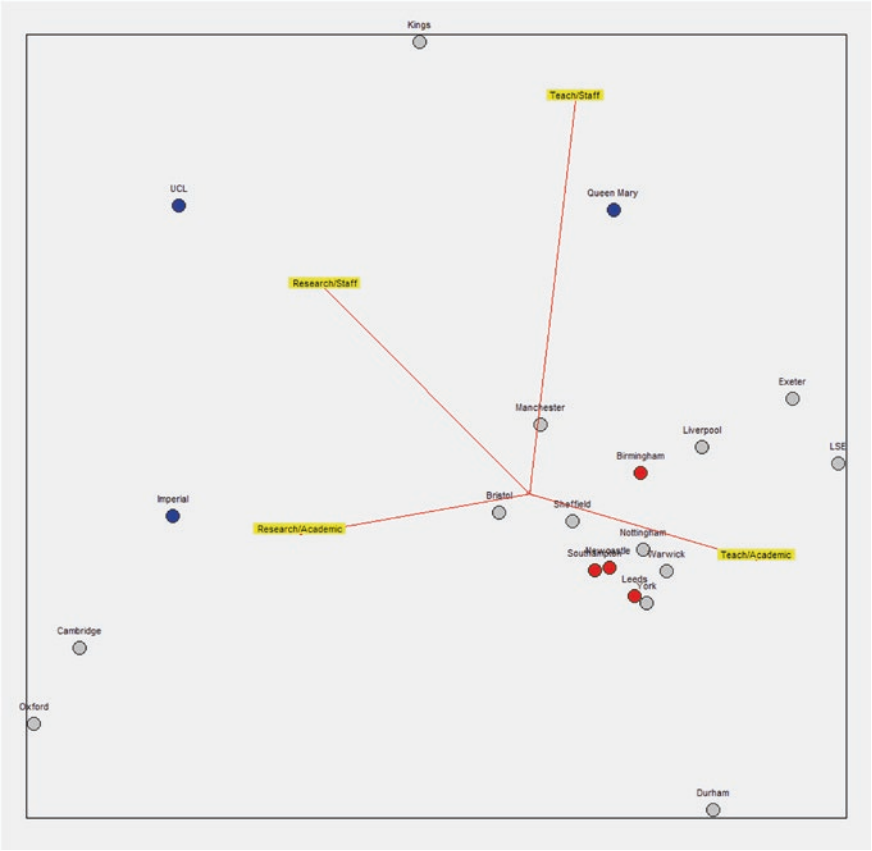

Fig. 1 Co-plot representation of the Higher Education Institutions

mapped by means of multidimensional scaling based on Guttman's smallest space analysis (1968). Finally, k arrows are drawn on the Euclidean space obtained in Stage 4. Each ratio j is represented by an arrow emerging from the center of gravity of the points from stage 3. The arrows associated with highly correlated information point in about the same direction. As a result, the cosines of angles between these arrows are approximately proportional to the correlations between their associated criteria. In our case, universities located high up on an arrow are likely to produce high output and utilize low input with respect to the specific ratio.

In Fig. 1, the four arrows represent each of the outputs divided by each of the inputs (income type divided by staff type). Each circle represents a university and the color has been chosen exogenously to the Co-plot model. Blue represents the universities that receive relatively consistent high ranks,

red represents those lowest in the rankings and gray are all the remaining universities. We see that Imperial and UCL are among the largest in the dataset and both receive high rankings due to their research income in comparison to their staff numbers. Queen Mary is far smaller in size and creates an umbrella for the more teaching-oriented universities in the dataset. Perhaps interestingly, Oxford and Cambridge are co-located in the bottom left of the plot but are not deemed as efficient or high up in the rankings as the combination of UCL, which earns higher income with fewer staff than Oxford, and Imperial, which earns lower income but with substantially lower staff levels than both Oxford and Cambridge.

11 Conclusions

The field of data envelopment analysis has grown at an exponential rate since the seminal papers of Farrell (1957) and Charnes et al. (1978). The original idea of evaluating after-school programs with multiple inputs and outputs has led to an enormous body of academic literature. Within this field is a sub-group of papers in which many researchers have sought to improve the differential capabilities of DEA and to fully rank both efficient, as well as inefficient, DMUs.

The DEA ranking concepts have been divided into seven general areas. The first group of papers is based on the super-efficiency approach, in which the efficient units may receive a score greater than one, as a function of how important the DMU is in pushing out the frontier. This idea has proved popular and many papers have applied the idea to multiple DEA variations, broadening its use from mere ranking to outlier detection, sensitivity analyses and scale classification. The issue of infeasibility appears to have been solved.

The second grouping is based on benchmarking, in which a DMU is highly ranked if it is chosen as a useful target for many other DMUs. This is of substantial use when looking to benchmark industries.

The third group of papers is based on a cross-efficiency matrix. By evaluating DMUs through both self and peer pressure, one can attain a potentially more balanced view of the DMU but at the cost of a loss of information.

The fourth group of papers creates an almost complete ranking based on comparing all DMUs through a production function approach with a common set of weights. The advantage of such a viewpoint is that it becomes possible to rank both efficient and inefficient units with respect to a single

hyper-frontier. The disadvantage is that it is less clear how to estimate the common set of weights and different approaches lead to completely different sets of rankings.

The fifth group of papers developed a connection between multivariate statistical techniques and DEA. Discriminant analysis is applied in order to compute a common set of weights, from which the DMUs can be ranked. In practice, non-parametric statistical tests showed a strong correlation between the final ranking and the original DEA dichotomous classification. In addition, combining principal component analysis and DEA reduces the bias caused by an excessive number of variables in comparison with observations, frequently leading to a ranking of DMUs without the need for additional preferential information from decision-makers.

In the sixth section, which crosses multi-criteria decision-making models with DEA, some concepts require additional, preferential information in order to aid the ranking process. The additional information can be incorporated into or alongside the standard DEA results through the use of assurance regions or discrimination intensity functions. Other concepts combined the two approaches without the need for additional information, such as the maximin efficiency ratio model and a multi-objective linear program. The most recent work on combining goal programming and DEA looks very promising.

The seventh and last group of papers discuss the ranking of inefficient units. One approach, entitled a measure of inefficiency dominance, ranks the inefficient units according to their average proportional inefficiency in all factors.

It should be noted that many papers have been written in an empirical context, utilizing the concepts discussed here. Our aim was to discuss the base methodologies rather than the subsequent applications, though it can be noted that certain techniques have been heavily used in specific areas. Super-efficiency has been applied in a wide range of papers from financial institutions and industry to public sector regulation, education and health care. Benchmarking has been used extensively in the field of utilities, industry and agricultural productivity. Cross-efficiency has been applied in many areas of manufacturing, including engineering design, flexible manufacturing systems, industrial robotics and business process re-engineering. It has also been used for project and R&D portfolio selection. The statistical techniques have been applied to universities and industry and MCDA/DEA to agriculture and the oil industry. Clearly, these methodologies have wide-ranging applicability in many areas of both the public and the private sectors.

Finally, many mathematical and statistical techniques have been presented here, all with the basic aim of increasing the discriminatory power of data envelopment analysis and ranking the DMU. However, while each technique may be useful in a specific area, no single methodology can be prescribed here as the panacea of all ills. It remains to be seen whether the ultimate DEA model can be developed to solve all problems and which will consequently be easy to solve by practitioners in the field and academics alike. It would seem more likely that specific models or combinations of models will be tailored to questions that arise over time.

Bibliography

Adler, N., and B. Golany. 2001. Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to Western Europe. *European Journal of Operational Research* 132 (2): 18–31.

Adler, N., and B. Golany. 2002. Including principal component weights to improve discrimination in data envelopment analysis. *Journal of the Operational Research Society* 53 (9): 985–991.

Adler, N., and A. Raveh. 2008. Presenting DEA graphically. *Omega* 36 (5): 715–729.

Adler, N., and E. Yazhemsky. 2010. Improving discrimination in data envelopment analysis: PCA–DEA or variable reduction. *European Journal of Operational Research* 202 (1): 273–284.

Adler, N., L. Friedman, and Z. Sinuany-Stern. 2002. Review of ranking methods in the data envelopment analysis context. *European Journal of Operational Research* 140 (2): 249–265.

Anand, S., and A. Sen. 1994. Human Development Index: Methodology and measurement. HDOCPA-1994-02, United Nations Development Programme (UNDP).

Andersen, P., and N.C. Petersen. 1993. A procedure for ranking efficient units in data envelopment analysis. *Management Science* 39 (10): 1261–1294.

Banker, R.D., and H. Chang. 2006. The super-efficiency procedure for outlier identification, not for ranking efficient units. *European Journal of Operational Research* 175 (2): 1311–1320.

Banker, R.D., A. Charnes, and W.W. Cooper. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science* 30 (9): 1078–1092.

Banker, R.D., H. Chang, and Z. Zheng. 2017. On the use of super-efficiency procedures for ranking efficient units and identifying outliers. *Annals of Operations Research* 250 (1): 21–35.

Bardhan, I., W.F. Bowlin, W.W. Cooper, and T. Sueyoshi. 1996. Models for efficiency dominance in data envelopment analysis. Part I: Additive models and MED measures. *Journal of the Operations Research Society of Japan* 39: 322–332.

Belton, V., and T.J. Stewart. 1999. DEA and MCDA: Competing or complementary approaches? In *Advances in Decision Analysis*, ed. N. Meskens and M. Roubens. Norwell: Kluwer Academic.

Charnes, A., W.W. Cooper, and E. Rhodes. 1978. Measuring the efficiency of decision-making units. *European Journal of Operational Research* 2: 429–444.

Charnes, A., C.T. Clark, W.W. Cooper, and B. Golany. 1985a. A developmental study of data envelopment analysis in measuring the efficiency of maintenance units in the US air forces. *Annals of Operations Research* 2: 95–112.

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

Charnes, A., W.W. Cooper, and S. Li. 1989. Using data envelopment analysis to evaluate the efficiency of economic performance by Chinese cities. *Socio-Economic Planning Science* 23: 325–344.

Charnes, A., W.W. Cooper, Z.M. Huang, and D.B. Sun. 1990. Polyhedral cone-ratio data envelopment analysis models with an illustrative application to large commercial banks. *Journal of Econometrics* 46: 73–91.

Chen, Y. 2004. Ranking efficient units in DEA. *Omega* 32 (3): 213–219.

Chen, Y. 2005. Measuring super-efficiency in DEA in the presence of infeasibility. *European Journal of Operational Research* 161 (2): 545–551.

Chen, Y., and L. Liang. 2011. Super-efficiency DEA in the presence of infeasibility: One model approach. *European Journal of Operational Research* 213: 359–360.

Chen, Y., J. Du, and J. Huo. 2013. Super-efficiency based on a modified directional distance function. *Omega* 41 (3): 621–625.

Cook, W.D., and D.A. Johnston. 1992. Evaluating suppliers of complex systems: A multiple criteria approach. *Journal of the Operations Research Society* 43: 1055–1061.

Cook, W.D., and M. Kress. 1990a. A data envelopment model for aggregating preference rankings. *Management Science* 36 (11): 1302–1310.

Cook, W.D., and M. Kress. 1990b. An m^{th} generation model for weak ranking of players in a tournament. *Journal of the Operations Research Society* 41 (12): 1111–1119.

Cook, W.D., and M. Kress. 1991. A multiple criteria decision model with ordinal preference data. *European Journal of Operational Research* 54: 191–198.

Cook, W.D., and M. Kress. 1994. A multiple-criteria composite index model for quantitative and qualitative data. *European Journal of Operational Research* 78: 367–379.

Cook, W.D., M. Kress, and L.M. Seiford. 1993. On the use of ordinal data in data envelopment analysis. *Journal of the Operations Research Society* 44: 133–140.

Cook, W.D., M. Kress, and L.M. Seiford. 1996. Data envelopment analysis in the presence of both quantitative and qualitative factors. *Journal of the Operations Research Society* 47: 945–953.

Cook, W.D., L. Liang, Y. Zha, and J. Zhu. 2009. A modified super-efficiency DEA model for infeasibility. *Journal of the Operational Research Society* 60 (2): 276–281.

Cooper, W.W., K.S. Park, and J.T. Pastor. 1999. RAM: A range adjusted measure of inefficiency for use with additive models, and relations to other models and measures in DEA. *Journal of Productivity Analysis* 11 (1): 5–42.

Cooper, W.W., J.T. Pastor, F. Borras, J. Aparicio, and D. Pastor. 2011. BAM: A bounded adjusted measure of efficiency for use with bounded additive models. *Journal of Productivity Analysis* 35 (2): 85–94.

de Carvalho Chaves, M.C., J.C.C. Soares de Mello, and L. Angulo-Meza. 2016. Studies of some duality properties in the Li and Reeves model. *Journal of the Operational Research Society* 67 (3): 474–482.

De Witte, K., and L. Hudrlikova. 2013. What about excellence in teaching? A benevolent ranking of universities. *Scientometric* 96: 337–364.

Doyle, J.R., and R. Green. 1994. Efficiency and cross-efficiency in data envelopment analysis: Derivatives, meanings and uses. *Journal of the Operations Research Society* 45 (5): 567–578.

Du, J., L. Liang, and J. Zhu. 2010. A slacks-based measure of super-efficiency in data envelopment analysis: A comment. *European Journal of Operational Research* 204 (3): 694–697.

Dula, J.H., and B.L. Hickman. 1997. Effects of excluding the column being scored from the DEA envelopment LP technology matrix. *Journal of the Operations Research Society* 48: 1001–1012.

Dyson, R.G., and E. Thanassoulis. 1988. Reducing weight flexibility in data envelopment analysis. *Journal of the Operations Research Society* 39: 563–576.

Farrell, M.J. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society* A120: 253–281.

Gandibleux, X. (ed.). 2006. *Multiple criteria optimization: State of the art annotated bibliographic surveys*, Vol. 52. New York: Springer Science & Business Media.

Ghasemi, M.R., J. Ignatius, and A. Emrouznejad. 2014. A bi-objective weighted model for improving the discrimination power in MCDEA. *European Journal of Operational Research* 233 (3): 640–650.

Golany, B. 1988. An interactive MOLP procedure for the extension of data envelopment analysis to effectiveness analysis. *Journal of the Operations Research Society* 39 (8): 725–734.

Golany, B., and Y.A. Roll. 1994. Incorporating standards via data envelopment analysis. In *Data envelopment analysis: Theory, methodology and applications*, ed. A. Charnes, W.W. Cooper, A.Y. Lewin, and L.M. Seiford. Norwell: Kluwer Academic.

Green, R.H., and J.R. Doyle. 1995. On maximizing discrimination in multiple criteria decision making. *Journal of the Operations Research Society* 46: 192–204.

Green, R.H., J.R. Doyle, and W.D. Cook. 1996. Preference voting and project ranking using data envelopment analysis and cross-evaluation. *European Journal of Operational Research* 90: 461–472.

Guo, I.L., H.S. Lee, and D. Lee. 2017. An integrated model for slack-based measure of super-efficiency in additive DEA. *Omega* 67: 160–167.

Guttman, L. 1968. A general non-metric technique for finding the smallest space for a configuration of points. *Psychometrica* 33: 469–506.

Halme, M., T. Joro, P. Korhonen, S. Salo, and J. Wallenius. 1999. A value efficiency approach to incorporating preference information in data envelopment analysis. *Management Science* 45 (1): 103–115.

Hashimoto, A. 1997. A ranked voting system using a DEA/AR exclusion model: A note. *European Journal of Operational Research* 97: 600–604.

Hazelkorn, E. 2015. *Rankings and the reshaping of higher education: The battle for world-class excellence*. New York: Springer.

Hwang, C.L., and K. Yoon. 1981. Methods for multiple attribute decision making. In *Multiple attribute decision making* (pp. 58–191). Berlin and Heidelberg: Springer.

Johnson, A.L., and L.F. McGinnis. 2009. The hyperbolic-oriented efficiency measure as a remedy to infeasibility of super efficiency models. *Journal of the Operational Research Society* 60 (11): 1511–1517.

Joro, T., P. Korhonen, and J. Wallenius. 1998. Structural comparison of data envelopment analysis and multiple objective linear programming. *Management Science* 44 (7): 962–970.

Kao, C., and H.T. Hung. 2005. Data envelopment analysis with common weights: the compromise solution approach. *Journal of the Operational Research Society* 56 (10): 1196–1203.

Kornbluth, J.S.H. 1991. Analyzing policy effectiveness using cone restricted data envelopment analysis. *Journal of the Operations Research Society* 42: 1097–1104.

Kritikos, M.N. 2017. A full ranking methodology in data envelopment analysis based on a set of dummy decision making units. *Expert Systems with Applications* 77: 211–225.

Lee, H.S., C.W. Chu, and J. Zhu. 2011. Super-efficiency DEA in the presence of infeasibility. *European Journal of Operational Research* 212 (1): 141–147.

Li, X.-B., and G.R. Reeves. 1999. A multiple criteria approach to data envelopment analysis. *European Journal of Operational Research* 115: 507–517.

Liang, L., J. Wu, W.D. Cook, and J. Zhu. 2008. Alternative secondary goals in DEA cross-efficiency evaluation. *International Journal of Production Economics* 113 (2): 1025–1030.

Lovell, C.K., and A.P.B. Rouse. 2003. Equivalent standard DEA models to provide super-efficiency scores. *Journal of the Operational Research Society* 54 (1): 101–108.

Oral, M., O. Kettani, and P. Lang. 1991. A methodology for collective evaluation and selection of industrial R&D projects. *Management Science* 7 (37): 871–883.

Raveh, A. 2000. Co-plot: A graphic display method for geometrical representations of MCDM. *European Journal of Operational Research* 125 (3): 670–678.

Ray, S.C. 2008. The directional distance function and measurement of super-efficiency: An application to airlines data. *Journal of the Operational Research Society* 59 (6): 788–797.

Rezaeiani, M.J., and A.A. Foroughi. 2018. Ranking efficient decision making units in data envelopment analysis based on reference frontier share. *European Journal of Operational Research* 264 (2): 665–674.

Roll, Y., W.D. Cook, and B. Golany. 1991. Controlling factor weights in data envelopment analysis. *IIE Transactions* 23 (1): 2–9.

Rubem, APdS, J.C.C.S. de Mello, and L.A. Meza. 2017. A goal programming approach to solve the multiple criteria DEA model. *European Journal of Operational Research* 260 (1): 134–139.

Ruiz, J.L., and I. Sirvent. 2016. *Ranking decision making units: The cross-efficiency evaluation. Handbook of operations analytics using data envelopment analysis*, ed. S.-N. Hwang, H.-S. Lee, and J. Zhu, 1–30. New York: Springer.

Seiford, L.M., and J. Zhu. 1999. Infeasibility of super-efficiency data envelopment analysis models. *Information Systems and Operational Research* 37 (2): 174–187.

Sexton, T.R., R.H. Silkman, and A.J. Hogan. 1986. Data envelopment analysis: Critique and extensions. In *Measuring efficiency: An assessment of data envelopment analysis*, ed. R.H. Silkman, 73–105. San Francisco: Jossey-Bass.

Sinuany-Stern, Z., and L. Friedman. 1998. Data envelopment analysis and the discriminant analysis of ratios for ranking units. *European Journal of Operational Research* 111: 470–478.

Sinuany-Stern, Z., A. Mehrez, and A. Barboy. 1994. Academic departments efficiency via data envelopment analysis. *Computers & Operations Research* 21 (5): 543–556.

Sinuany-Stern, Z., A. Mehrez, and Y. Hadad. 2000. An AHP/DEA methodology for ranking decision making units. *International Transactions in Operational Research* 7: 109–124.

Sueyoshi, T. 1999. Data envelopment analysis non-parametric ranking test and index measurement: Slack-adjusted DEA and an application to Japanese agriculture cooperatives. *Omega* 27: 315–326.

Thanassoulis, E., and R.G. Dyson. 1992. Estimating preferred target input-output levels using data envelopment analysis. *European Journal of Operational Research* 56: 80–97.

Thompson, R.G., E. Lee, and R.M. Thrall. 1992. DEA/AR efficiency of U.S. independent oil/gas producers over time. *Computers and Operations Research* 19 (5): 377–391.

Thompson, R.G., F.D. Singleton, R.M. Thrall, and B.A. Smith. 1986. Comparative site evaluations for locating a high-energy physics lab in Texas. *Interfaces* 16: 35–49.

Thompson, R.G., L.N. Langemeier, C.T. Lee, and R.M. Thrall. 1990. The role of multiplier bounds in efficiency analysis with application to Kansas farming. *Journal of Econometrics* 46: 93–108.

Thrall, R.M. 1996. Duality, classification and slacks in data envelopment analysis. *Annals of Operations Research* 66: 109–138.

Tone, K. 2001. A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research* 130 (3): 498–509.

Torgersen, A.M., F.R. Forsund, and S.A.C. Kittelsen. 1996. Slack-adjusted efficiency measures and ranking of efficient units. *The Journal of Productivity Analysis* 7: 379–398.

Troutt, M.D. 1995. A maximum decisional efficiency estimation principle. *Management Science* 41: 76–82.

Wong, Y.-H.B., and J.E. Beasley. 1990. Restricting weight flexibility in data envelopment analysis. *Journal of the Operations Research Society* 41: 829–835.

Xue, M., and P.T. Harker. 2002. Note: Ranking DMUs with infeasible super-efficiency DEA models. *Management Science* 48 (5): 705–710.

Young, F.W., and R.M. Hamer. 1987. *Multidimensional scaling: History, theory and applications*. London: Lawrence Erlbaum.

Zhu, J. 1996a. Robustness of the efficient decision-making units in data envelopment analysis. *European Journal of Operational Research* 90: 451–460.

Zhu, J. 1996b. Data envelopment analysis with preference structure. *Journal of the Operations Research Society* 47: 136–150.

Zhu, J. 2000. Multi-factor performance measure model with an application to Fortune 500 companies. *European Journal of Operational Research* 123 (1): 105–124.

Distributional Forms in Stochastic Frontier Analysis

Alexander D. Stead, Phill Wheat and William H. Greene

1 Introduction

This chapter provides an overview of a particular aspect of stochastic frontier analysis (SFA). The SF model is typically used to estimate best-practice 'frontier' functions that explain production or cost and predict firm efficiency relative to these. Extensive reviews of the broad stochastic frontier (SF) methodology are undertaken by Kumbhakar and Lovell (2000), Murillo-Zamorano (2004), Coelli et al. (2005), Greene (2008), and Parmeter and Kumbhakar (2014). This review will focus on the many different uses of various distributional forms.

Section 2 begins with a brief account of the motivation and development of efficiency analysis and prediction based on the standard SF model. A key feature of SF models is the focus on unobserved disturbance in the econometric model. This entails a deconvolution of the disturbance into a firm inefficiency component— quantification of which is the goal of the analysis—and a statistical noise term. Following this general outline, we discuss approaches to dealing with some key specification issues. Section 3 considers alternative distributional assumptions for inefficiency. Section 4 examines panel data issues.

A. D. Stead · P. Wheat
University of Leeds, Leeds, UK

W. H. Greene (✉)
Stern School of Business, New York University, New York, NY, USA
e-mail: wgreene@stern.nyu.edu

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_8

Section 5 considers modelling heteroskedasticity in error terms and its usefulness for policy analysis. Section 6 considers alternative noise distributions within SF models. Section 7 considers amendments to the standard SF model when the data contains efficient firms. Section 8 considers other received proposals relevant to appropriate distributional assumptions in SFA. Section 9 concludes.

2 Departure Points

The standard theory of the firm holds that firms seek to maximise profit. Under certain assumptions, a *profit function* exists that reflects the maximum profit attainable by the firm. The profit function is derived from the firm's *cost function*, which represents the minimum cost given outputs and input prices, and its *production function*, which describes the firm's technology. These are 'frontier' functions in the sense that they represent optimal outcomes that firms cannot improve upon given their existing technology. The duality of the production and cost functions was demonstrated by Shephard (1953). Debreu (1951) introduced the notion of a distance function to describe a multiple output technology and proposed that the radial distance of a producer's outputs from the distance function be used as a measure of technical inefficiency. Koopmans (1951) provided a definition of technical efficiency.

The idea that firms might depart from profit maximisation was first suggested in passing by Hicks (1935), who speculated that firms with market power in particular may choose to enjoy some of their rents not as profit, but as reduced effort to maximise profits, or 'a quiet life'. Later, Leibenstein (1966, 1975) discussed various empirical indications of firm-level 'X-inefficiency' and how it might arise. The debate between Leibenstein (1978) and Stigler (1976) highlighted two alternative characterisations of inefficiency: as a result of selective rationality and non-maximising behaviour, resulting in non-allocative welfare loss, or as the redistribution of rents within the firm, and therefore consistent with the idea of maximising outward behaviour. The latter characterisation essentially posits that firms are maximising an objective function including factors other than profit, and encompasses a wide range of specific hypotheses about firm behaviour. The revenue maximisation hypothesis of Baumol (1967), the balanced growth maximisation hypothesis of Marris (1964) and the expense preference hypothesis of Williamson (1963) are examples of hypotheses within which the firm (or its managers, given informational asymmetry between principal

and agent) pursues other objectives jointly with profit or subject to a profit constraint. We should therefore bear in mind that when we discuss efficiency, it is relative to an objective that we define, and not necessarily that of the firm (or its agents).

The early literature on econometric estimation of cost functions has focal points at Johnston (1960) for the UK coal industry and Nerlove (1963) for US electricity generation. These authors focused primarily on estimation of the shape of the empirical cost or production functions. Typically, ordinary least squares (OLS) was used to estimate a linear model:

$$y_i = x_i\beta + \varepsilon_i, \tag{1}$$

where y_i is cost or output, β is a vector of parameters to be estimated, ε_i is a random error term, $i = 1, 2, \ldots, I$ denotes an observed sample of data and x_i is the vector of independent variables. In the case of a production function, independent variables include input quantities and other factors affecting production, while in the case of a cost frontier, independent variables include output quantities and input prices, along with other factors affecting cost (Shephard 1953). Commonly, the dependent and independent variables are logged, in order to linearise what is assumed to be a multiplicative functional form. Note that the estimation of (1) via least squares, where a symmetric error term is assumed, is only consistent with the idea of a frontier function if we assume that firms are all fully efficient, and that departures from the estimated frontier are explained purely by measurement error and other random factors, such as luck. This fact has motivated many alternative proposals that are consistent with the notion of a frontier. Farrell (1957) proposed the use of linear programming to construct, assuming constant returns to scale, a piecewise linear isoquant and to define technical inefficiency as the radial distance of the firm from this isoquant.

An approach that amends (1) so the error is one-sided, yields a deterministic or 'full' frontier specification, in which the residuals are attributed entirely to inefficiency. Since a firm must be operating on or below its production frontier, and on or above its cost frontier, this means that $s\varepsilon_i \leq 0$, where $s = 1$ for a production frontier and $s = -1$ for a cost frontier. Aigner and Chu (1968) suggested linear or quadratic programming approaches to deterministic frontier estimation. Respectively, these minimise $\sum_{i=1}^{I} \varepsilon_i$ or $\sum_{i=1}^{I} \varepsilon_i^2$, subject to the constraint that $s\varepsilon_i \leq 0$. Schmidt (1976) noted that these are maximum likelihood (ML) estimators under the assumptions that the error term is exponentially or half normally distributed. Omitting the restriction that the residuals be one-sided leads to OLS and least absolute deviations (LAD) estimation, which would be ML estimation under the

assumptions that ε_i follows the normal or Laplace distributions, two-sided counterparts of the half-normal and exponential distributions, respectively. Afriat (1972) proposed a deterministic frontier model in which $\exp(\varepsilon_i)$ follows a two-parameter beta distribution, to be estimated via ML, which as Richmond (1974) noted is equivalent to assuming a gamma distribution for ε_i. The usual regularity conditions for ML estimation do not hold for deterministic frontier functions, since the range of variation of the dependent variable depends upon the parameters. Greene (1980) points out that under certain specific assumptions, this irregularity is actually not the relevant constraint. Specifically, if both the density and first derivative of the density of ε converge to zero at the origin, then the log-likelihood function is regular for ML estimation purposes. Deterministic frontier models with gamma and log-normal error term are examples.

Deterministic frontier models suffer from a serious conceptual weakness. They do not account for noise caused by random factors such as measurement error or luck. A firm whose production is impacted by a natural disaster might by construction appear to be inefficient. In order to account for measurement error, Timmer (1971) suggested amending the method so that the constraint $s\varepsilon_i \leq 0$ holds only with a given probability, thereby allowing a proportion of firms to lie above (below) the production (cost) frontier. However, this probability must be specified in advance in an arbitrary fashion. An alternative proposal made by Aigner et al. (1976) has the error drawn from a normal distribution with variance $\sigma^2\theta$ when $s\varepsilon_i \leq 0$ and $\sigma^2(1-\theta)$ when $s\varepsilon_i > 0$, where $0 < \theta < 1$. Essentially, though this is not made explicit, this allows for normally distributed noise with variance $\sigma^2(1-\theta)$ and inefficiency implicitly following a half-normal distribution with variance $(1-2/\pi)\sigma^2(1-\theta)$, under the assumption that where $s\varepsilon_i \leq 0$ firms are fully efficient. The resulting likelihood function is that of a 50:50 mixture of two differently scaled normal distributions truncated at zero from the left and right, respectively. The discontinuity of this specification once again violates the standard regularity conditions for ML estimation.

The issues with the models suggested by Timmer (1971) and Aigner et al. (1976) stem in both cases from their peculiar assumption that firms must be fully efficient when $s\varepsilon_i \leq 0$, which remains rooted in an essentially deterministic view of frontier estimation. The current literature on SFA, which overcomes these issues, begins with Aigner et al. (1977) and Meeusen and van Den Broeck (1977). They proposed a composed error:

$$\varepsilon_i = v_i - su_i \tag{2}$$

where v_i is a normally distributed noise term with zero mean, capturing random factors such as measurement error and luck, and u_i is a non-negative random variable capturing inefficiency and is drawn from a one-sided distribution. Battese and Corra (1977) proposed an alternative parameterisation of the model. Given particular distributional assumptions about the two error components, the marginal distribution of the composed error ε_i may be derived by marginalising u_i out of the joint probability;

$$f_\varepsilon(\varepsilon_i) = \int_0^\infty f_v(\varepsilon_i + su_i) f_u(u_i) \mathrm{d}u_i \tag{3}$$

where f_ε, f_v, and f_u are the density functions for ε_i, v_i, and u_i, respectively. The half-normal and exponential distributions were originally proposed for u_i. Assuming a normal distribution for v_i, the resulting distributions for ε_i are the skew-normal distribution, studied by Weinstein (1964) and Azzalini (1985), and the exponentially modified Gaussian distribution originally derived by Grushka (1972).

The ultimate objective of SFA is deconvolution of estimated residuals into separate predictions for the noise and inefficiency components. The latter is the focus of efficiency analysis. Since the parameters of f_u are outputs of the estimation process, we obtain an estimated distribution of efficiency, and as proposed by Lee and Tyler (1978), the first moment of this estimated distribution may be used to predict overall average efficiency. However, decomposing estimated residuals into observation-specific noise and efficiency estimates was elusive until Jondrow et al. (1982) suggested predicting based on the conditional distribution of $u_i|\varepsilon_i$, which is given by

$$f_{u|\varepsilon}(u_i|\varepsilon_i) = \frac{f_v(\varepsilon_i + su_i) f_u(u_i)}{f_\varepsilon(\varepsilon_i)}. \tag{4}$$

They derived (4) for the normal-half normal and normal-exponential cases. The mean, $E(u_i|\varepsilon_i)$, and mode, $M(u_i|\varepsilon_i)$, of this distribution were proposed as predictors. Waldman (1984) examined the performance of these and other computable predictors. Battese and Coelli (1988) suggest the use of $E\left[\exp(-u_i)|\varepsilon_i\right]$ when the frontier is log-linear. Kumbhakar and Lovell (2000) suggest that this is more accurate than $\exp[-E(u_i|\varepsilon_i)]$, especially when u_i is large. In practice, the difference often tends to be very small. It should be noted that the distribution of the efficiency predictions, $E(u_i|\varepsilon_i)$ will not match the unconditional, marginal distribution of

the true, unobserved u_i. Wang and Schmidt (2009) derived the distribution of $E(u_i|\varepsilon_i)$ and show that it is a shrinkage of u_i towards $E(u_i)$, with $E(u_i|\varepsilon_i) - u_i$ approaching zero as $\sigma_v^2 \to 0$.

3 Alternative Inefficiency Distributions

The efficiency predictions of the stochastic frontier model are sensitive to the assumed distribution of u_i. A number of alternatives have been proposed. Several two-parameter generalisations of the half-normal and exponential distributions, respectively, allow for greater flexibility in the shape of the inefficiency distribution, with non-zero modes in particular. The flexible forms generally enable testing against their simpler nested distributions. Stevenson (1980) proposed the truncated normal model[1]; Greene (1990) and Stevenson (1980) proposed gamma distributions. The truncated normal distribution, denoted $N^+\left(\mu, \sigma_u^2\right)$, nests the half normal when its location parameter μ (the pre-truncation mean) is zero, and its mode is μ when $\mu \geq 0$. The similar 'folded normal distribution' denoted $\left|N\left(\mu, \sigma_u^2\right)\right|$, i.e. that of the absolute value of an $N\left(\mu, \sigma_u^2\right)$ normal random variable, also nests the half normal when μ is zero, but has a non-zero mode only when $\mu \geq \sigma_u$ (Tsagris et al. 2014; Hajargasht 2014).

The gamma distribution with shape parameter k and scale parameter σ_u nests the exponential distribution when $k = 1$. A two-parameter lognormal distribution, which resembles the gamma distribution, for u_i is adopted by Migon and Medici (2001). It is possible to adopt even more flexible distributional assumptions; Lee (1983) proposed using a very general four-parameter Pearson distribution for u_i as a means of nesting several simpler distributions. On the other hand, Hajargasht (2015) proposed a one-parameter Rayleigh distribution for u_i which has the attraction of being a parsimonious way of allowing for a non-zero mode. Griffin and Steel (2008) proposed a three-parameter extension of Greene's two-parameter gamma model that nests the gamma, exponential, half-normal and (heretofore never considered) Weibull models. Some of these represent minor extensions of the base case models. In all cases, however, the motivation is a more flexible, perhaps less restrictive characterisation of the variation of efficiency across

[1]In the SF literature, 'truncated normal' refers specifically to the left truncation at zero of a normal distribution with mean μ and variance σ_u^2.

firms. In many cases, the more general formulations nest more restrictive, but common distributional forms.

The inefficiency distributions discussed above were proposed to enable more flexible distributional assumptions about u_i. Other proposals have addressed specific practical and theoretical issues. One is the 'wrong skew' problem, which is discussed in more detail below. Broadly, the skewness of su_i should be negative, both in the theory and as estimated using data. In estimation, it often happens that the information extracted from the data suggests skewness in the wrong direction. This would seem to conflict with the central assumption of the stochastic frontier model. The problem for the theoretical specification is that, since $\text{Skew}(\varepsilon_i) = \text{Skew}(v_i) - s\text{Skew}(u_i) = -s\text{Skew}(u_i)$ when v_i is symmetrically distributed, the skewness of the composed error ε_i is determined by that of u_i. Therefore, imposing $\text{Skew}(u_i) > 0$ implies that $-s\text{Skew}(\varepsilon_i) > 0$. Since all of the aforementioned distributions for u_i allow only for positive skewness, this means that the resulting SF models cannot handle skewness in the 'wrong' direction. An estimated model based on sample data will typically give an estimate of zero for $\text{Var}(u_i)$ if the estimated skewness (however obtained) goes in the wrong direction.

'Wrong skew' could be viewed as a finite sample issue, as demonstrated by Simar and Wilson (2010). Even when the assumed distribution of ε_i is correct, samples drawn from this distribution can have skewness in the 'wrong' direction with some probability that decreases with the sample size. Alternatively, it may indeed be the case that, though non-negative, the distribution of u_i has a zero or negative skew, and therefore, our distributional assumptions need to be changed accordingly. To this end, Li (1996) and Lee and Lee (2014)[2] consider a uniform distribution, $u_i \sim U(a, b)$, so that u_i and ε_i are both symmetric, and Carree (2002) and Tsionas (2007) consider the binomial distribution and Weibull distributions, respectively, which both allow for skewness in either direction. Arguably, these 'solutions' are ad hoc remedies to what might be a fundamental conflict between the data and the theory. Notwithstanding the availability of these remedies, negative skewness, defined appropriately is a central feature of the model.

Also relevant here are SF models with 'bounded inefficiency'. These are motivated by the idea that there is an upper bound on inefficiency beyond

[2]Lee and Lee (2014) focus on the upper bound on inefficiency in the normal-uniform model and appear to have been unaware of the model's earlier introduction by Li (1996), who was motivated by the skewness issue.

which firms cannot survive. Such a boundary could be due to competitive pressure, as suggested by Qian and Sickles (2008). However, we also consider that it could arise in monopolistic infrastructure industries which are subject to economic regulation, since depending on the strength of the regulatory regime, some inefficiency is likely to be tolerated.[3]

Implementation of bounded inefficiency involves the *right-truncation* of one of the canonical inefficiency distributions found in the SF literature. The upper tail truncation point is a parameter that would be freely estimated and is interpreted as the inefficiency bound. Lee (1996) proposed a tail-truncated half-normal distribution for inefficiency, and Qian and Sickles (2008) and Almanidis and Sickles (2012) propose a more general 'doubly truncated normal' distribution (i.e. the tail truncation of a truncated normal distribution). Almanidis et al. (2014) discuss the tail-truncated half-normal, tail-truncated exponential and doubly truncated normal inefficiency distributions. The latter of these may have positive or negative skewness depending on its parameter values. In fact, it is clear that this may be true of the right-truncation of many other non-negative distributions with non-zero mode.

A difficulty with certain distributional assumptions is that the integral in (3) may not have a closed-form solution, so that there may not be an analytical expression for the log-likelihood function. This issue first arose in the SF literature in the case of the normal-gamma model, in which case the problem was addressed in several different ways. Stevenson (1980) noted that relatively straightforward closed-form expressions exist for integer values of the shape parameter k, of the normal-gamma model and derived the marginal density of ε_i for $k = 0$, $k = 1$, and $k = 2$. Restricting k to integer values gives the Erlang distribution, so this proposal amounts to a restrictive normal-Erlang model. The need to derive distinct formulae for every possible integer value of k makes this approach unattractive. Beckers and Hammond (1987) derived a complete log-likelihood for the normal-gamma model, but due to its complexity their approach has not been implemented. Greene (1990) approximated the integral using quadrature, but this approximation proved rather crude (Ritter and Simar 1997). An alternative approach, proposed by Greene (2003), is to approximate the integral via simulation

[3]Such 'tolerance' does not necessarily reflect the technical competence or experience of regulators per se. It could reflect the perceived limitations on the robustness of the analysis (e.g. data quality), which necessitates a risk averse efficiency finding from a regulatory review.

in order to arrive at a maximum simulated likelihood (MSL) solution. For more detail on MSL estimation, see Train (2009). In the context of SFA, Greene and Misra (2003) note that the simulation approach could be used to approximate the integral in (3) for many distributional assumptions as long as the marginal variable u_i can be simulated. Since the integral is the expectation of $f_v(\varepsilon_i + su_i)$ given the assumed distribution for u_i, it can be approximated by averaging over Q draws from the distribution of u_i:

$$f_\varepsilon(\varepsilon_i) = \int_0^\infty f_v(\varepsilon_i + su_i) f_u(u_i) \mathrm{d}u_i \approx \frac{1}{Q} \sum_{q=1}^{Q} f_v\left[\varepsilon_i + sF_u^{-1}(d_q)\right] \tag{5}$$

where d_q is draw number q from the standard uniform distribution, transformed by the quantile function F_u^{-1} into a draw from the distribution of u_i. In cases in which there is no analytical F_u^{-1}, such as the normal-gamma model, the integral may nevertheless be expressed in terms of an expectation that may be approximated via simulation. Greene (2003) recommends using Halton sequences, which aim for good coverage of the unit interval, rather than random draws from the uniform distribution, in order to reduce the number of draws needed for a reasonable approximation of the integral.

As an alternative to simulation, various numerical quadrature approaches may be used. Numerical quadrature involves approximating an integral by a weighted sum of values of the integrand at various points. In many cases, this involves partitioning the integration interval and approximating the area under the curve within each of the resulting subintervals using some interpolating function. The advantage of quadrature over simulation lies in speed of computation, given that the latter's time-consuming need to obtain potentially large numbers of independent draws for each observation. However, it may be challenging to find appropriate quadrature rules in many cases. Another alternative, proposed by Tsionas (2012), is to approximate f_ε using the (inverse) fast Fourier transform of the characteristic function of f_ε. The characteristic function, φ_ε, is the Fourier transform of f_ε, and as shown by Lévy's inversion theorem (see Theorem 1.5.4 in Lukacs and Laha 1964), the inverse Fourier transform of the characteristic function can be used to obtain f_ε. Since the Fourier transform of a convolution of two functions is simply the product of their Fourier transforms, i.e. $\varphi_\varepsilon = \varphi_v \varphi_u$ (see Bracewell 1978, p. 110), φ_ε may be relatively simple even when f_ε has no closed form, and f_ε may be approximated by the inverse fast Fourier transform of φ_ε. On the basis of Monte Carlo experiments, Tsionas (2012) finds that this is a faster method for approximating f_ε in the normal-gamma

and normal-beta cases than either Gaussian quadrature or Monte Carlo simulation, with the former requiring a large number of quadrature points and the latter an even larger number of draws for comparable accuracy. This approach has not yet been adopted as widely as simulation, perhaps due to its relative complexity.

A natural question would be which, of the many alternatives discussed above, is the most appropriate distribution for inefficiency? Unfortunately, theory provides little guidance on this question. Oikawa (2016) argues that a simple Bayesian learning-by-doing model such as that of Jovanovic and Nyarko (1996), in which a firm (or manager) maximises technical efficiency given prior beliefs about and previous realisations of an unknown technology parameter, supports a gamma distribution for inefficiency. However, Tsionas (2017) shows that this conclusion is sensitive to the sampling of, and assumed prior for, the firm-specific parameter, and that under alternative formulations there is no basis for favouring the gamma distribution (or any known distribution). Furthermore, both authors assume that firms maximise expected profits, whereas alternative behavioural assumptions may yield very different results. Of course, the choice of inefficiency distribution may be driven by practical considerations, such as a need to allow for wrong skewness or to estimate an upper bound on inefficiency. The question of which inefficiency distribution to use is an empirical one and leads us to consider testing in the context of SFA. As noted previously, some of the more flexible inefficiency distributions nest simpler distributions. In these cases, we may test against to simpler nested models. For example, we may test down from the normal-gamma to the normal-exponential model by testing the null hypothesis that $k = 1$. We may test down from the normal-truncated normal (or the normal-folded normal) to the normal-half normal model by testing the null hypothesis that $\mu = 0$. These are standard problems.

There are some remaining complications in the specification search for the SF model. We may wish to test for the presence of the one-sided error, often interpreted as a test for the presence of inefficiency. In this case, the errors are normally distributed under the null hypothesis $H_0 : \sigma_u = 0$. This is a non-standard problem because the scale parameter σ_u is at a boundary of the parameter space under H_0. Case 5 in Self and Liang (1987) shows that where a single parameter of interest lies on the boundary of the parameter space under the null hypothesis, the likelihood ratio (LR) statistic follows a 50:50 mixture of χ^2_0, and χ^2_1 distributions, denoted $\chi^2_{1:0}$, for which the 95% value is 2.706 (Critical values are presented in Kodde and Palm 1986). Lee (1993) finds that this is the case under $H_0 : \sigma_u = 0$ in the normal-half

normal model. A Lagrange multiplier test for this case in the SF model is developed in Lee and Chesher (1986).

This result does not apply when f_u has two or more parameters. Coelli (1995) states that, in the normal-truncated normal model, the LR statistic under $H_0 : \sigma_u = \mu = 0$ follows a 25:50:25 mixture of χ_0^2, χ_1^2 and χ_2^2 distributions, and that this is a special case of the result for two restrictions in Gouriéroux et al. (1982), which deals with inequality restrictions.[4] This result matches Case 7 in Self and Liang (1987), in which two parameters of interest lie on the boundary of the parameter space under the null. The test seems to have been incorrectly applied; under $H_0 : \sigma_u = \mu = 0$, only one parameter lies on the boundary. Equivalently, viewing the test as a one-tailed test of $H_0 : \sigma_u \leq 0, \mu = 0$, we only have one inequality restriction. Case 6 in Self and Liang (1987), in which there are two parameters of interest, one on a boundary, and one not on a boundary, seems to be more applicable, suggesting a 50:50 mixture of χ_1^2 and χ_2^2 distributions, denoted $\chi_{2:1}^2$. More fundamentally, $H_0 : \sigma_u = \mu = 0$ may not be the appropriate null hypothesis: when the scale parameter of the inefficiency distribution is set to zero, all other parameters of the distribution are in fact unidentified. Equivalently, a normal distribution for ε_i can be recovered in the normal-truncated normal case as $\mu \to -\infty$, for any value of σ_u. The general problem of testing when there are unidentified nuisance parameters under the null hypothesis is discussed by Andrews (1993a, b) and Hansen (1996). To our knowledge has not been addressed in the SF literature.

We may wish to choose between two non-nested distributions. In this case, Wang et al. (2011) suggest testing goodness of fit by comparing the distribution of the estimated residuals to the theoretical distribution of the compound error term. This is a simpler method than, for example, comparing the distribution of the efficiency predictions to the theoretical distribution of $E(u|\varepsilon)$ as derived by Wang and Schmidt (2009), since the distribution of the compound error is much simpler. For example, as discussed previously, ε_i follows a skew-normal distribution in the normal-half normal model, and an exponentially modified Gaussian distribution in the normal-exponential model. Under alternative specifications, the distribution of the estimated residuals may become rather complex, however.

[4]If we view the normal-half normal model as a skew-normal regression model in which we expect (but do not restrict) the skewness parameter σ_u/σ_v to be positive, then we view the test for the presence of inefficiency as a one-tailed test of the H_0 that $\sigma_u \leq 0$, or equivalently that $\sigma_u/\sigma_v = 0$, rather than as a test involving a boundary issue. Comparing the case of one inequality constraint in Gouriéroux et al. (1982) to Case 5 in Self and Liang (1987), we see the same result.

4 Panel Data

The basic panel data SF model in the contemporary literature is as in (1) with the addition of a t subscript to denote the added time dimension of the data:

$$y_{it} = x_{it}\beta + \varepsilon_{it}, \tag{6}$$

where $t = 1, 2, \ldots, T$. The composite error term is now

$$\varepsilon_{it} = \alpha_i + v_{it} - su_{it}. \tag{7}$$

Along with the usual advantages of panel data, Schmidt and Sickles (1984) identify three benefits specific to the context of SFA. First, under the assumption that inefficiency is either time invariant or that it varies in a deterministic way, efficiency prediction is consistent as $T \to \infty$. In contrast, this is not the case as $N \to \infty$. Second, distributional assumptions can be rendered less important, or avoided altogether, in certain panel data specifications. In particular skewness in the residual distribution does not have to be the only defining factor of inefficiency. Instead, time persistence in inefficiency can be exploited to identify it from random noise. Third, it becomes possible, using a fixed-effects approach, to allow for correlation between inefficiency and the variables in the frontier.[5] In addition, the use of panel data allows for the modelling of dynamic effects.

In the context of panel data SF modelling, one of the main issues is the assumption made about the variation (or lack thereof) of inefficiency over time. Another is the way in which we control (or do not control) for firm-specific unobserved heterogeneity and distinguishes this from inefficiency. For the purposes of this discussion, we divide the received panel data SF models into three classes: models in which inefficiency is assumed to be time-invariant, models in which inefficiency is time-varying, and models which control for unobserved heterogeneity with either time-invariant or time-varying inefficiency. To finish this section, we consider briefly multi-level panel datasets and the opportunities that they provide for analysis.

[5]However, since Schmidt and Sickles (1984), cross-sectional models have been proposed, such as those of Kumbhakar et al. (1991), Huang and Liu (1994), and Battese and Coelli (1995), that allow for dependence between inefficiency and frontier variables. These are discussed in Sect. 4.

4.1 Time-invariant Efficiency

One approach to panel data SFA is to assume that efficiency varies between firms but does not change over time, as first proposed by Pitt and Lee (1981). Referring to (6) and (7), the basic panel data SF model with time-invariant efficiency assumes that $\alpha_i = 0, u_{it} = u_i$, so that we have:

$$y_{it} = x_{it}\beta + v_{it} - su_i. \tag{8}$$

This specification has the advantage that prediction (or estimation) of u_i is consistent as $T \to \infty$. The appeal of this result is diminished given that the assumption of time-invariance is increasingly hard to justify as the length of the panel increases. In contrast to the cross-sectional case, there is no need to assume that u_i is a random variable with a particular distribution, and therefore, there are several different methods may be used to estimate (8), depending on our assumptions about u_i.

Schmidt and Sickles (1984) proposed four alternative approaches. First, we may assume that u_i is a firm-specific fixed effect, and to estimate the model using either a least squares dummy variable (LSDV) approach, in which u_i is obtained as the estimated parameter on the dummy variable for firm i, or equivalently by applying the within transformation, in which case u_i is obtained as firm i's mean residual. Second, we may assume that u_i is a firm-specific random effect and estimate the model using feasible generalised least squares (FGLS). The difference between the fixed-effects and random-effects approaches is that the latter assumes that the firm-specific effects are uncorrelated with the regressors, while the former does not. Third, Schmidt and Sickles (1984) suggested instrumental variable (IV) estimation of the error components model proposed by Hausman and Taylor (1981) and Amemiya and MaCurdy (1986), which allows for the firm-specific effect to be correlated with some of the regressors and uncorrelated with others, and is thus intermediate between the fixed-effects and random-effects models. Fourth, as in Pitt and Lee (1981), u_i could be regarded as an independent random variable with a given distribution, as in the cross-sectional setting, with the model being estimated via ML.

The first three approaches share the advantage that no specific distributional assumption about u_i is required. As a consequence, the estimated firm-specific effects could be positive. As a result, firm-specific efficiency can only be measured relative to the best in the sample, not to an absolute benchmark. The estimated u_i is given by

$$u_i = \max_j sa_j - sa_i, \tag{9}$$

where a_i is the estimated firm-specific effect for firm i. The fixed-effects specification has the advantage of allowing for correlation between u_i and x_{it}. But the drawback is that time-invariant regressors cannot be included, meaning that efficiency estimates will be contaminated by any differences due to time-invariant variables. The assumption that the factors are uncorrelated with errors (noise or inefficiency) can be examined using the Hausman test (Hausman 1978; Hausman and Taylor 1981). If this assumption appears to hold, a random effects approach such as Pitt and Lee (1981) may be preferred. Another approach is to estimate a correlated random-effects model using Chamberlain-Mundlak variables—see Mundlak (1978) and Chamberlain (1984)—to allow for correlation between the random effects and the regressors. Griffiths and Hajargasht (2016) propose correlated random effects SF models using Chamberlain-Mundlak variables to allow for correlation between regressors and error components, including inefficiency terms.

The ML approach to estimation of (8) was first suggested by Pitt and Lee (1981), who derived an SF model for balanced panel data with a half-normal distribution for u_i and a normal distribution for v_{it}. This model therefore nests the basic cross-sectional model of Aigner et al. (1977) when $T = 1$. As in the cross-sectional setting, alternative distributional assumptions may be made. Battese and Coelli (1988) generalise the Pitt and Lee (1981) model in two ways: first, by allowing for an unbalanced panel and second, by assuming a truncated normal distribution for u_i. Normal-exponential, normal-gamma and normal-Rayleigh variants of the Pitt and Lee (1981) model are implemented in LIMDEP Version 11 (Greene 2016). As in the cross-sectional setting, parameter estimates and efficiency predictions obtained under the ML approach are more efficient than those from semi-parametric models if the distributional assumptions made are valid. If those assumptions are not valid, they may be inconsistent and biased. To be sure, the ability to test distributional assumptions is very limited.

4.2 Time-Varying Efficiency

Allowing for variation in efficiency over time is attractive for a number of reasons. As already noted, the assumption that efficiency is time-invariant is increasingly hard to justify as T increases. We would expect average efficiency to change over time. There may also be changes in the relative positions of firms, in terms of convergence or divergence in efficiency between firms, and potentially also changes in rankings through firms overtaking

each other. A wide variety of time-varying efficiency SF specifications have been proposed, each differing with respect to their flexibility in modelling the time path of efficiency and each having their own advantages and disadvantages.

As Amsler et al. (2014) note, panel data SF specifications can be grouped into four categories with respect to how u_{it} changes over time. One of these, covered in the preceding section, is models with time-invariant efficiency, so that $u_{it} = u_i$. Second, we could assume independence of u_{it} over t. In this case, we may simply estimate a pooled cross-sectional SF model, the possibility of unobserved heterogeneity notwithstanding. The advantages of this approach are the flexibility of u_{it}—and by extension, that of $E(u_{it}|\varepsilon_{it})$—over time, its simplicity and its sparsity, given that it adds no additional parameters to the model. However, the assumption of independence over time is clearly inappropriate.

Third, we may treat u_{it} as varying deterministically over time. One approach is to include time-varying fixed or random effects, a_{it}, with u_{it} being given by

$$u_{it} = \max_j sa_{jt} - sa_{it}. \tag{10}$$

Of course, given that $N \leq IT$ firm- and time-specific parameters[6] cannot be identified, some structure must be imposed. Kumbhakar (1991, 1993) proposed combining firm-specific (but time-invariant) and time-specific (but firm-invariant) effects, such that $a_{it} = \lambda_i + \sum_{t=2}^{T} \lambda_t$. This imposes a common trend in u_{it} among firms, albeit one that may be quite erratic. Lee and Schmidt (1993) proposed a specification, $a_{it} = \lambda_t \alpha_i$, which again imposes a trend over time. This is common for all firms, but complicates estimation due to its non-linearity. An alternative approach is to specify that $a_{it} = g(t)$ as proposed by Cornwell et al. (1990), who specifically suggested a quadratic time trend with firm-specific parameters, such that $a_{it} = \lambda_i + \lambda_{i1} t + \lambda_{i2} t^2$. This specification is flexible, in that it allows for firms to converge, diverge or change rankings in terms of efficiency. Ahn et al. (2007) propose a specification which nests both the Lee and Schmidt (1993) and Cornwell et al. (1990) models, in which $a_{it} = \sum_{j=1}^{p} \lambda_{jt} \alpha_{ji}$, thus allowing for arbitrary, firm-specific time trends. This specification nests the Lee and Schmidt (1993) model when $p = 1$, and the Cornwell et al. (1990) model when $p = 3$, $\lambda_{1t} = 1$, $\lambda_{2t} = t$, $\lambda_{3t} = t^2$. The value of p is estimated along with

[6] N being the total number of observations, so that $N = IT$ in the case of a balanced panel.

the model parameters. The authors discuss estimation and identification of model parameters. Ahn et al. (2013) discuss estimation of this model when there are observable variables correlated with the firm-specific effects, but not with v_{it}. An alternative approach based on factor modelling and allowing for arbitrary, smooth, firm-specific efficiency trends is proposed by Kneip et al. (2012).

Because semi-parametric models yield only relative estimates of efficiency, it is not possible to disentangle the effects of technical change (movement of the frontier) and efficiency change. An analogous approach in the context of parametric specifications is to use a 'scaling function', so that

$$u_{it} = g(t)u_i. \tag{11}$$

Here, u_i is a time-invariant random variable following a one-sided distribution—as in the time-invariant specification of Pitt and Lee (1981)—and $g(t)$ is a non-negative function of t. Kumbhakar (1990) proposed $g(t) = 1/\left[1 + \exp\left(\lambda_1 t + \lambda_2 t^2\right)\right]$; Battese and Coelli (1992) proposed $g(t) = \exp\left[\lambda_1(t - T)\right]$ and $g(t) = \exp\left[\lambda_1(t - T) + \lambda_2(t - T)^2\right]$. In each case, u_i is assumed to follow half-normal distribution. In these models, efficiency moves in the same direction for all firms, but there may be convergence of firms over time. In addition, with the exception of the one-parameter Battese and Coelli (1992) scaling function, these allow for non-monotonic trends in u_{it} over time. However, they do not allow for changes in rank over time, which requires firm-specific time trends.

Cuesta (2000) generalised the one-parameter Battese and Coelli (1992) scaling function to allow for firm-specific time trends, so that $g(t) = \exp\left[\lambda_{1i}(t - T)\right]$. An extension to the two-parameter case would be straightforward. This allows for firm-specific time trends, as in the Cornwell et al. (1990) model, but again at the cost of increasing the number of parameters in the model by a factor of I. However, Wheat and Smith (2012) show that the Cuesta (2000) specification, unlike that of Battese and Coelli (1992), can lead to a counterintuitive 'falling off' of firms with high $E(u_{it}|\varepsilon_{it})$ in the final year of the sample They propose a model in which $g(t) = \exp\left[\lambda_{1i}(t - \lambda_{2i})\right]$, that does not have the same feature.[7] More generally, as Wheat and Smith (2012) note, the many different models using that use functions are sensitive to the precise form of $g(t)$ in terms of parameter estimates, fit and efficiency predictions.

[7]Clearly, this model is far from parsimonious, since $g(t)$ includes $2I$ parameters. In fact, the authors apply a simpler model, $g(t) = \exp\left[\lambda_{1i}(t - \lambda_2)\right]$ after failing to reject $H_0 : \lambda_{2i} = \lambda_2$.

A fourth approach to time-variation of u_{it} in panel data SF models is to allow for correlation between u_{it} over time by assuming that $(u_{i1}, \ldots, u_{iT})$ are drawn from an appropriate multivariate distribution. Among their various proposals, Pitt and Lee (1981) suggested that $(u_{i1}, \ldots, u_{iT})$ could be drawn from a multivariate truncated normal distribution. They abandoned this approach, after noting that the likelihood function for this model involves intractable T-dimensional integrals.[8] In addition, Horrace (2005) showed that the marginal distribution of u_{it} in this case is not truncated normal. However, as suggested by Amsler et al. (2014), it is possible to specify a multivariate distribution with the desired marginal distributions, and also obviate T-dimensional integration when evaluating $\ln L$, by using a copula function. Sklar's theorem—see Nelsen (2006, pp. 17–14)—states that any multivariate cumulative density function can be expressed in terms of a set of marginal cumulative density functions and a copula. For example, we have

$$H_u(u_{i1}, \ldots, u_{iT}) = C[F_{u1}(u_{i1}), \ldots, F_{uT}(u_{iT})] \tag{12}$$

where H_u is a multivariate cumulative density function for $(u_{i1}, \ldots, u_{iT})$, $C[.]$ is the copula function, and $F_{u1}(u_{i1}), \ldots, F_{uT}(u_{iT})$ are the marginal cumulative density functions for u_{it} for each time period. We would normally assume that $F_{ut} = F_u$ for all t, so that we have

$$H_u(u_{i1}, \ldots, u_{iT}) = C[F_u(u_{i1}), \ldots, F_u(u_{iT})]. \tag{13}$$

From this, it can be seen that the probability density function is given by

$$h_u(u_{i1}, \ldots, u_{iT}) = \prod_{t=1}^{T} \left[f_u(u_{it})\right] c[F_u(u_{i1}), \ldots, F_u(u_{iT})] \tag{14}$$

where c is the derivative of the copula. It follows from this that a multivariate density $h(u_{i1}, \ldots, u_{iT})$ with the desired marginal densities given by f_u can be obtained by combining f_u and F_u with an appropriate copula density c. Many different copula functions exist—it is beyond the scope of this chapter to review the various candidates—each embodying different dependence structures. Note that $c = 1$ relates to the special case of independence. This allows marginal distributions to be specified, but the

[8]The authors instead estimate a system of T equations via the seemingly unrelated regressions (SUR) model proposed by Zellner (1962). However, this approach offers no way of predicting observation-specific efficiencies.

problem of T-dimensional integration to evaluate the log-likelihood persists. For this reason, Amsler et al. (2014) propose and implement an alternative approach whereby instead of specifying a copula for $(u_{i1}, \ldots, u_{iT})$, a copula is specified for the composite errors $(\varepsilon_{i1}, \ldots, \varepsilon_{iT})$. In this case, we have

$$h_{\varepsilon}(\varepsilon_{i1}, \ldots, \varepsilon_{iT}) = \prod_{t=1}^{T} \left[f_{\varepsilon}(\varepsilon_{it})\right] c[F_{\varepsilon}(\varepsilon_{i1}), \ldots, F_{\varepsilon}(\varepsilon_{iT})] \tag{15}$$

where h_{ε} is the multivariate distribution for $(\varepsilon_{i1}, \ldots, \varepsilon_{iT})$ and F_{ε} is the marginal cumulative density function for ε_{it}. In this case, an appropriate marginal distribution for ε_{it} is chosen, such as the skew-normal distribution. In this case, the correlation is between the composite errors, introducing dependency between both error components. Amsler et al. (2014) take both approaches, estimating a model in which $(\varepsilon_{i1}, \ldots, \varepsilon_{iT})$ is drawn from a joint distribution as in (15) via ML, and a model in which $(u_{i1}, \ldots, u_{iT})$ is drawn from a joint distribution as in (14), while v_{it} is assumed independent, via MSL. A Gaussian copula function is used in both cases. The authors discuss prediction of efficiency. In this case, it is based on $u_{it}|\varepsilon_{i1}, \ldots, \varepsilon_{iT}$. This results in improved predictions relative to those based on $u_{it}|\varepsilon_{it}$, since the composite errors from all years are informative about u_{it} when there is dependency between them.

The copula approach proposed by Amsler et al. (2014) is attractive, since it can be seen as intermediate between the pooled SF approach and the approach of specifying SF models with deterministically time-varying u_{it}. As such, it retains the advantage of the latter approach in allowing for dependency over time, without specifying a particular functional form for the time trend. It also obviates the large number of additional parameters otherwise needed to allow flexibility with respect to time trends. Rather than some factor of I, the number of new parameters is limited to the correlation coefficients $\rho_{ts} \forall t \neq s$. A number of simplifying assumptions can be made to reduce the number of these while retaining flexibility. Firms may converge or diverge, or change rankings, using a relatively parsimonious specification under this approach.

4.3 Unobserved Heterogeneity

Aside from considerations of the appropriate way to model trends in u_{it} over time, which is peculiar to the panel data SF context, more general panel data issues are also relevant. Primary among these is the need to account

for possible unobserved heterogeneity between firms. In general, this means incorporating firm-specific effects which are time-invariant but not captured by the regressors included in the frontier. These may be either correlated or uncorrelated with the regressors, i.e. they may be fixed or random effects, respectively. In general, failure to account for fixed effects may bias parameter estimates, while failure to account for random effects generally will not.[9] In the SF context, failure to account for fixed or random effects means such effects may be attributed to u_{it}.

A number of models have been proposed which incorporate fixed or random effects. These are interpreted as capturing unobserved heterogeneity rather than as inefficiency effects. Kumbhakar (1991) proposed extending the pooled cross-section model to incorporate firm and time effects uncorrelated with the regressors, so that

$$\varepsilon_{it} = v_{it} + a_i + a_t - su_{it} \tag{16}$$

where a_i and a_t are firm- and time-specific fixed or random effects. In the fixed-effects case, Kumbhakar (1991) suggests estimation via ML with firm dummy variables, under the assumptions that a_i, a_t, and v_{it} are drawn from normal distributions with zero means and constant variances, and u_{it} is drawn from a truncated normal distribution. A simplified version of this model, omitting a_t and treating a_i as a fixed effect, was used by Heshmati and Kumbhakar (1994). This model was also considered by Greene (2004, 2005a, b), who proposed the specification

$$\varepsilon_{it} = v_{it} + a_i - su_{it} \tag{17}$$

where a_i is a time-invariant fixed or random effect, and the specification is referred to as the 'true fixed effects' (TFE) or 'true random effects' (TRE) model, accordingly. In the TFE case, estimation proceeds by simply replacing the constant term in the standard pooled with a full set of firm dummies and estimating the model via ML. However, evidence presented by Greene (2005b) from Monte Carlo experiments suggests that this approach suffers from the incidental parameters problem. As a result, Chen et al. (2014) propose an alternative ML approach based on the within transformation, which is not subject to this problem, and Belotti and Ilardi (2018) extend this approach to allow for heteroscedastic u_{it}.

[9]However, in the context of a log-linear model, the estimate of the intercept will be biased in either case.

In the TRE case, Greene (2004, 2005a, b) proposed estimation of the model via MSL, assuming that $a_i \sim N\left(0, \sigma_a^2\right)$. Greene (2005b) notes that the TRE approach—and indeed the standard SF model—can be seen as special cases of a random parameters model and proposes a random parameters specification incorporating heterogeneity in β, so that

$$y_{it} - x_i\beta_i = \varepsilon_{it} = v_{it} - su_{it} \tag{18}$$

where β_i is assumed to follow a multivariate normal distribution with mean vector β and covariance matrix $\sum$. The random intercept is $\beta_{0i} = \beta_0 + a_i$ in terms of the TRE notation. The model is estimated via MSL. The resemblance of this approach to the Bayesian SF specifications considered by Tsionas (2002) is noted. However, the Bayesian approach has the drawback of requiring some prior distribution to be chosen for all parameters, including those of f_u. Greene (2008) notes that in the classical framework, 'randomness' of the parameters reflects technological heterogeneity between firms, whereas in the Bayesian framework, 'randomness' of the parameters is supposed to reflect the uncertainty of the analyst.[10]

A discrete approximation to the random parameters SF model is possible using a latent class approach to capture heterogeneity in some or all of the β parameters, as proposed by Orea and Kumbhakar (2004). In this specification, each firm belongs to one of J classes, each class having a distinct technology, so that for class j, we have technology parameters β_j. Class membership is unknown. Each firm is treated as belonging to class j with unconditional probability p_j, where the unconditional probabilities are estimated as parameters after normalising such that $\sum_{j=1}^{J} p_j = 1$ (leaving $J-1$ additional parameters to be estimated). The model may be estimated via ML. Conditional probabilities of class membership for each observation obtained by

$$p_{ij} = \frac{p_j f_\varepsilon\left(y_{it} - x_i\beta_j\right)}{\sum_{j=1}^{J}\left[p_j f_\varepsilon\left(y_{it} - x_i\beta_j\right)\right]} \tag{19}$$

The primary issue with the TFE and TRE and similar models is that any time-invariant effects are attributed to a_i, when it is entirely possible that they should, partly or wholly, be attributed to u_{it}. Several recent proposals therefore extend this modelling approach to allow for u_{it} to be broken down

[10]Despite this, Tsionas (2002) does interpret the models as incorporating technological heterogeneity.

into separate time-invariant and time-varying components capturing 'persistent' and 'transient' inefficiency effects, respectively. Thus,

$$u_{it} = w_i + w_{it} \tag{20}$$

where typically both w_i and w_{it} are random variables drawn from some one-sided distribution. A similar decomposition of u_{it} was first suggested by Kumbhakar and Heshmati (1995), who proposed that $u_{it} = a_i + w_{it}$ and Kumbhakar and Hjalmarsson (1995), who proposed $u_{it} = a_i + \alpha_t + w_{it}$, where a_i and α_t are firm- and time-specific fixed or random effects, respectively.[11] Colombi et al. (2014) and Tsionas and Kumbhakar (2014) propose an extension of the TRE model, accordingly referred to as the generalised true random effects (GTRE) model, in which

$$\varepsilon_{it} = v_{it} + a_i - s(w_i + w_{it}) \tag{21}$$

This model therefore includes four error components, allowing for noise, unobserved heterogeneity, and persistent and transient inefficiency. Identification requires specific distributional assumptions to be made about either a_i or w_i, or both. The following distributional assumptions are typically made: $v_{it} \sim N\left(0, \sigma_v^2\right)$, $a_i \sim N\left(0, \sigma_a^2\right)$ and w_i and w_{it} are follow half-normal distributions with constant variances. Each of the error components is assumed to be independent. Various approaches to estimation of the GTRE model have been proposed. Kumbhakar et al. (2014) suggest a multi-step approach. In the first step, a standard random effects panel data model including a noise component $v_{it}^* = v_{it} + w_{it}$ and a time-invariant random effects component $a_i^* = a_i + w_i$. This can be estimated via FGLS, avoiding any explicit distributional assumptions. Subsequently, the estimates of these error components are used as the dependent variables in separate constant-only SF models, which decompose them into their two-sided and one-sided components. This is straightforward to implement using standard software packages.

Alternatively, Colombi et al. (2014) use the result that ε_{it} in the GTRE model is the sum of two random variables, each drawn from an independent closed skew-normal distribution.[12] As its name suggests, the closed

[11]Note that these proposals are very similar to those of Kumbhakar (1991) and Heshmati and Kumbhakar (1994), the difference being the interpretation of a_i and α_t as picking up inefficiency effects, rather than unobserved heterogeneity.

[12]The univariate skew normal distribution is a special case of the closed skew-normal distribution. To see that ε_{it} is the sum of two closed skew-normal random variables, therefore, consider that $v_{it}^* = v_{it} + w_{it}$. and $a_i^* = a_i + w_i$ both follow skew-normal distributions. For details on the closed skew-normal distribution, see González-Farías et al. (2004).

skew-normal distribution is closed under summation—see Proposition 2.5.1 of González-Farías et al. (2004b) or Theorem 1 in González-Farías et al. (2004a). Therefore, ε_{it} follows a skew-normal distribution. This enables estimation of the model via ML. However, Filippini and Greene (2016) note that this is extremely challenging, since the log-likelihood involves the probability density function for a T-variate normal distribution and the cumulative density function for a $T + 1$-variate normal distribution. They proposed a simpler approach based on MSL, which exploits the fact that the GTRE model is simply the TRE model in which the time-invariant error component follows a skew-normal distribution. Colombi et al. (2014) show how to obtain predictions for w_i and w_{it}.

The attraction of the GTRE model is that it is quite general, in that it allows for the decomposition of the composite error into noise, random effects, persistent inefficiency, and transient inefficiency components. It also nests various simpler models, such as the TRE model, the standard pooled SF model, the Pitt and Lee (1981) model, and a standard random-effects model. However, Badunenko and Kumbhakar (2016) recently concluded on the basis of Monte Carlo experiments that the model is very limited in its ability to precisely predict the individual error components in practice, and suggest that the model may not outperform simpler models in many cases.

4.4 Multi-level Panel Datasets

This section has outlined how panel data allows for a richer characterisation of efficiency and thus panel data is desirable for undertaking efficiency analysis. Both Smith and Wheat (2012) and Brorsen and Kim (2013) have considered using data on a number of organisations over time, but disaggregated on sub-firm divisions (henceforth: plants) for each organisation. Thus, there are two levels of data

$$y_{its} = x_{ijt}\beta + \tau_{its} + v_{its} \tag{22}$$

which is (6) but with the addition of a plant subscript j. There are two key advantages to considering data of this form. Firstly, such an approach allows for the measurement of internal efficiency variation within an organisation, as well as simultaneously measuring efficiency against comparator organisations (external efficiency). Smith and Wheat (2012) propose a model (ignoring the time dimension for simplicity) in which $u_{ij} = u_i + u_{ij}^*$, where u_i is a one-sided component common to all of firm i's plants, and u_{ij}^* is a plant-specific component assumed to follow a half-normal distribution. The

authors suggest estimating the model using a two-step approach, in which u_i is obtained from a fixed or random effect in the first step. Note that, in the one-period or pooled cross-section cases, this is simply the panel data specification of Kumbhakar and Hjalmarsson (1995) and Kumbhakar and Heshmati (1995).

Lai and Huang (2013) argue that there is likely to be intra-firm correlation between both plant-level efficiency and noise effects. Rather than allow for separate correlations between the v_{ij} and the u_{ij}, the authors propose a model in which the ε_{ij} are correlated such that $\rho\left(\varepsilon_{ij}, \varepsilon_{il}\right) = \rho$. The components and v_{ij} and u_{ij} are assumed to be drawn from marginal normal and half-normal distributions, respectively, the authors allow for correlation between the composed errors using a Gaussian copula.

Secondly, both Brorsen and Kim (2013) and Smith and Wheat (2012) demonstrate that there is a need to model costs at the level that management autonomy resides. Failure to do so can result in misleading predictions of efficiency as it mismatches returns to scale properties of the cost function with efficiency. Brorsen and Kim (2013) used data on schools and school districts to show that if the model were estimated using data at district level then returns to scale are found to be decreasing rather than finding that these schools are inefficient. Ultimately, the aggregation bias is resulting in correlation between errors and regressors, since true measures of scale/density (at the disaggregate level) are not included in the model.

5 Heteroscedasticity and Modelling Inefficiency

In many applications of SFA, the analyst is interested in not only in the estimation or prediction of efficiency, but also in its variation in terms of a set of observable variables. However, the standard SF model assumes that u_i is independent of observed variables. Many applications, including Pitt and Lee (1981) as an early example, take a two-step approach to modelling efficiency: first, a standard SF model is estimated and used to generate efficiency predictions, and second, these predictions are regressed on a vector of explanatory variables. However, the second-step regression violates the assumption of independence in the first step, and Wang and Schmidt (2002) show that the two-step approach is severely biased. Given that u_i is a random variable, appropriate approaches involve specifying one or more parameters of the error distributions as a function of a set of covariates.

Deprins and Simar (1989a, b), Reifschneider and Stevenson (1991), Kumbhakar et al. (1991), Huang and Liu (1994), and Battese and Coelli (1995) all propose extensions of the basic SF model whereby

$$u_i = g(z_i, \delta) + w_i \tag{23}$$

where z_i is a vector of 'environmental' variables influencing inefficiency, δ is a vector of coefficients, and w_i is a random error. In the Deprins and Simar (1989a, b) specification, $g(z_i, \delta) = \exp(z_i\delta)$ and $w_i = 0$, and the model may estimated via non-linear least squares or via ML assuming $v_i \sim N\left(0, \sigma_v^2\right)$.[13] Reifschneider and Stevenson (1991) propose restricting both components of u_i to be non-negative, i.e. $g(z_i, \delta), w_i \geq 0$, though as Kumbhakar and Lovell (2000) and Greene (2008) note, this is not required for $u_i \geq 0$. An alternative approach was proposed by Kumbhakar et al. (1991), in which $g(z_i, \delta) = 0$ and w_i is the truncation at zero of a normally distributed variable with mean $z_i\delta$ and variance σ_u^2. Huang and Liu (1994) proposed a model in which $g(z_i, \delta) = z_i\delta$ and w_i is the truncation at $-z_i\delta$ of an $N\left(0, \sigma_u^2\right)$ random variable. The latter two models are in fact equivalent, as noted by and Battese and Coelli (1995). In simple terms, the model assumes that $v_i \sim N\left(0, \sigma_v^2\right)$ and $u \sim N^+\left(\mu_i, \sigma_u^2\right)$, where $\mu_i = z_i\delta$. Note that a constant term is included in z_i, so that the model nests the normal-truncated normal model of Stevenson (1980) and the normal-half normal model.

Another set of models, motivated by the desire to allow for heteroskedasticity in u_i, specify the scale parameter, rather than the location parameter, of the distribution of u_i as a function of a set of covariates.[14] Reifschneider and Stevenson (1991) first proposed amending the normal-half normal model so that $\sigma_{ui} = h(z_i), h(z_i) \in (0, \infty)$, but did not make any particular suggestions about $h(z_i)$ other than noting that the function must be constrained to be non-negative. Caudill and Ford (1993) suggested the functional form $\sigma_{ui} = \sigma_u(z_i\gamma)^\alpha$, which nests the standard homoskedastic normal-half normal model when $\alpha = 0$. Caudill et al. (1995) suggested a slightly sparser specification in which $\sigma_{ui} = \sigma_u \exp(z_i\gamma)$, and Hadri (1999) proposed a similar 'doubly heteroskedastic' SF model, $\sigma_{vi} = \exp(z_i\theta), \sigma_{ui} = \exp(z_i\gamma)$.

[13]Note that the authors in fact proposed a deterministic frontier model in which $E(u_i|z_i) = \exp(z_i\delta)$, but if we interpret the random error as v_i rather than a component of u_i, we have an SF model with a deterministic u_i.

[14]Note, however, that since the (post-truncation) variance of the truncated normal distribution is a function of the pre-truncation mean, the Kumbhakar et al. (1991), Huang and Liu (1994), and Battese and Coelli (1995) model also implies heteroskedasticity in u_i.

The approaches discussed above can be combined for an encompassing model in which both the location and scale parameters are functions of z_i. Wang (2002) proposed a model in which $u_i \sim N^+\left(\mu_i, \sigma_{ui}^2\right)$, where $\mu_i = z_i\delta$ and $\sigma_{ui}^2 = \exp(z_i\gamma)$, while Kumbhakar and Sun (2013) took this a step further, estimating a model in which $u_i \sim N^+\left(\mu_i, \sigma_{ui}^2\right)$ and $v_i \sim N\left(0, \sigma_{vi}^2\right)$, where $\mu_i = z_i\delta$, $\sigma_{vi} = \exp(z_i\theta)$, and $\sigma_{ui} = \exp(z_i\gamma)$, effectively combining the Hadri (1999) 'doubly heteroskedastic' model with that of Kumbhakar et al. (1991), Huang and Liu (1994), and Battese and Coelli (1995).[15]

Given the motivation of explaining efficiency in terms of z_i, and since z_i enters the model in a non-linear way. It is desirable to calculate the marginal effect of these z_{li}, the lth environmental variable, on efficiency. Of course, given that u_i is a random variable, we can only predict the marginal effect of z_l on *predicted* efficiency, and this means that the marginal effects formula used depends fundamentally on the efficiency predictor adopted. Where $u_i \sim N^+\left(\mu_i, \sigma_{ui}^2\right), \mu_i = z_i\delta$, the parameter δ_l is the marginal effect of z_{li} on the mode of the distribution of u_i, except when $z_i\delta \leq 0$. The derivative of the unconditional mode predictor,

$$\partial M(u_i)/\partial z_{li} = \begin{cases} \delta_l, z_i\delta > 0 \\ 0, z_i\delta \leq 0 \end{cases}. \tag{24}$$

Therefore, the unconditional mode yields a relatively simple marginal effect. Alternatively, Wang (2002) derived a marginal effects formula based on the derivative of the unconditional mean, $\partial E(u_i)/\partial z_{li}$. As the author shows, since $E(u_i)$ depends on the scale parameter, as well as the location parameter, of the distribution, marginal effects calculated using this formula can be non-monotonic even if z_{li} enters both functions in a linear fashion. This lends itself to potentially useful discussion of the 'optimal' (i.e. efficiency maximising) level of z_{li}. As noted by Hadri (1999), the variables entering μ_i, σ_{vi}, and σ_{ui} need not be the same in practice.

The efficiency prediction is usually based on the distribution of $u_i|\varepsilon_i$ (specifically its mean) rather than u_i. Kumbhakar and Sun (2013) argue that marginal effects should be based on $\partial E(u_i|\varepsilon_i)/\partial z_{li}$ rather than $\partial E(u_i)/\partial z_l$. and show that in this case, marginal effects depend upon the parameters not only of f_u but also of f_v and upon ε_i, i.e. all of the model's variables and parameters. Stead (2017) derives a marginal effects formula based on the conditional mode, $\partial M(u_i|\varepsilon_i)/\partial z_{li}$, which is somewhat

[15]Note the two similar but subtly different parameterisations, $\sigma_{ui} = \exp(z_i\gamma)$ and $\sigma_{ui}^2 = \exp(z_i\gamma)$.

simpler, particularly when both $\sigma_{vi} = \sigma_v$ and $\sigma_{ui} = \sigma_u$ in which case $\partial M(u_i|\varepsilon_i)/\partial z_{li} = \delta_l\left[\sigma_v^2/\left(\sigma_v^2 + \sigma_u^2\right)\right]$ when $M(u_i|\varepsilon_i) > 0$. Note that the marginal effects formulae discussed so far relate to changes in predicted u_i rather than predicted efficiency: Stead (2017) derives a marginal effect based on the Battese and Coelli (1988) predictor, $\partial E\left[\exp(-u_i)|\varepsilon_i\right]/\partial z_{li}$,, and notes that other formulae should be transformed into inefficiency space by multiplying by $-\exp\left(-\hat{u}_i\right)$ where $\hat{u}_i$ is the predictor for u_i since $\partial \exp\left(-\hat{u}_i\right)/\partial z_{li} = -(\partial \hat{u}_i/\partial z_{li})\exp\left(-\hat{u}_i\right)$. The choice between conditional and unconditional marginal effects formulae is between prediction of marginal effects for specific observations, and quantifying the relationship between environmental variables and inefficiency in general.

The idea that marginal effects should be based on a predictor of $u_i|\varepsilon_i$ rather than u_i has the appeal that the marginal effects discussed are consistent with the preferred efficiency predictor, in the sense that they indicate the change in predicted efficiency resulting from a change in z_{li}. On the other hand, such marginal effects are sensitive to changes in the frontier variables and parameters and the parameters of f_v, despite the fact that efficiency is not specified in this way. Another drawback is that while $\partial E(u_i)/\partial z_{li}$ and $\partial M(u_i)/\partial z_{li}$ are parameters for which standard errors and confidence intervals may be estimated, $\partial E(u_i|\varepsilon_i)/\partial z_{li}$ and $\partial M(u_i|\varepsilon_i)/\partial z_{li}$ are random variables for which prediction intervals are the only appropriate estimate of uncertainty, making hypothesis testing impossible. Kumbhakar and Sun (2013) suggest a bootstrapping approach to derive confidence intervals for $\partial E(u_i|\varepsilon_i)/\partial z_{li}$, but this is inappropriate since it treats ε_i as known.[16]

Given the rather complex marginal effects implied by the models discussed above, alternative specifications with simpler marginal effects have been proposed. Simar et al. (1994) propose that z_i should enter as a scaling function, such that $u_i = f(z_i\eta)u_i^*$, where u_i^* is assumed to follow some non-negative distribution that does not depend on z_i, and $f(z_i\eta)$ is a non-negative scaling function similar to those used in Battese and Coelli (1992) type panel data models. Wang and Schmidt (2002) note several features of this formulation: first, the shape of the distribution of u_i is the same for all observations, with $f(z_i\eta)$ simply scaling the distribution; models with this property are described as having the 'scaling property'. Second, it may yield relatively simple marginal effects expressions, e.g. when

[16]Note the similarity of the issues here to those around 'confidence intervals' and prediction intervals for $E(u_i|\varepsilon_i)$, discussed by Wheat et al. (2014).

$f(z_i\eta) = \exp(z_i\eta)$ or similar.[17] Third, as suggested by Simar et al. (1994), the β and η may be estimated via non-linear least squares without specifying a particular distribution for u_i^*. The scaling property is discussed further by Alvarez et al. (2006), who suggested testing for the scaling property.

More recently, Amsler et al. (2015) suggested an alternative parameterisation such that z_i enters the model through the post-truncation, rather than the pre-truncation, parameters of f_u. For example, the left truncation at zero of an $N\left(\mu_i, \sigma_{ui}^2\right)$ random variable, which we have denoted $N^+\left(\mu_i, \sigma_{ui}^2\right)$, may be reparameterised in terms of $E(u_i)$ and $\text{VAR}(u_i)$; that is, f_u may be expressed in terms of these parameters, and as a result, so may f_ε. The authors show that marginal effects are simpler and easier to interpret when environmental variables enter the model such that $E(u_i) = g(z_i, \delta), \text{VAR}(u_i) = h(z_i, \gamma)$ than when $\mu_i, \sigma_{ui}^2 = g(z_i, \delta), \sigma_{ui}^2 = h(z_i, \gamma)$. This is intuitive, given that we predict based on post-truncation parameters of f_u or $f_{u|\varepsilon}$. This approach is complicated somewhat by the requirement that $E(u_i) > \text{VAR}(u_i)$, as shown by Eq. (3) in Barrow and Cohen (1954), Eq. (16) in Bera and Sharma (1999), and Lemma 1 of Horrace (2015). For this reason, the authors suggest a specification in which $\text{VAR}(u_i) = \exp(z_i\gamma)$ and $E(u_i) = \text{VAR}(u_i) + \exp(z_i, \delta)$.

An additional motivation for the models discussed in this section is the analysis of production risk. Bera and Sharma (1999) proposed, in the context of a production frontier model, that $\text{VAR}(u_i|\varepsilon_i)$ be used as a measure of 'production uncertainty or risk. Note however that this is a far more restrictive measure than that used in the wider literature on production risk, which is variability of output, measured, for example, by $\text{VAR}(y_i)$. Nevertheless, these models offer considerable flexibility in modelling production risk according to this definition. Just and Pope (1978) showed that a drawback of log-linear (non-frontier) production function specifications, in which $q_i = \exp(y_i)$, is that the marginal production risk (i.e. the partial derivative of production risk) with respect to a given variable must always be the same as that variable's marginal product. The authors proposed an alternative specification with an additive error term multiplied by a scaling function. The form allows for variables that affect production and production risk in potentially opposite directions for variables that affect one but not the other. Kumbhakar (1993) and Battese et al. (1997) proposed SF variants

[17]However, the authors' discussion overstates the simplicity of marginal effects in this case, since it focuses on $\partial \ln \hat{u}_i / \partial z_{li}$, which is η_l regardless of the distribution of u_i^* (or indeed the choice of predictor). However, $\partial \hat{u}_i / \partial z_{li}$ is more complex, and as previously noted, the translation into efficiency space via $\partial \exp\left(-\hat{u}_i\right) / \partial z_{li}$ adds additional complexity.

of this model by including an inefficiency term u_i. Note, however, that any SF model in which one or both error terms are heteroskedastic allows for observation-specific production risk.

6 Alternative Noise Distributions

In the standard SF model, the noise term is assumed to follow a normal distribution. In contrast to the many different proposals concerning the distribution of u_i, discussed in Sect. 3, the distribution of v_i has received relatively little attention. This is perhaps natural, given that the main focus of SFA is on estimation or prediction of the former component. Nevertheless, consideration of alternative distributions for v_i is important for at least two main reasons. First, the standard model is not robust to outliers caused by noise, i.e. when the true noise distribution has thick tails. Second, and perhaps more importantly, the distribution of v_i has implications for the deconvolution of ε_i into noise and inefficiency components. Specifically, the distribution of $u_i|\varepsilon_i$, on which efficiency prediction is typically based, is influenced by f_v as well as f_u, as shown in (4).

The latter point in particular is not trivial. A change in distributional assumptions regarding v_i affects the degree of shrinkage of u_i towards $E(u_i)$ using $E(u_i)$.[18] A change in the assumed noise distribution can even be sufficient to change the rankings of firms[19] by altering the monotonicity properties of $E(u_i|\varepsilon_i)$ with respect to ε_i, which are in turn linked to the log-concavity properties of f_v. Ondrich and Ruggiero (2001) prove that $E(u_i|\varepsilon_i)$ is a weakly (strictly) monotonic function of ε_i for any weakly (strictly) log-concave f_v. Since the normal density is strictly log-concave everywhere, $E(u_i|\varepsilon_i)$ is a monotonic function of ε_i in the standard model. Under alternative noise distributions for which f_v is not strictly log-concave everywhere, there may be a weakly monotonic or even non-monotonic relationship between $E(u_i|\varepsilon_i)$ and ε_i. Such relationships have been noted in several studies proposing alternative, heavy tailed, noise distributions, which are discussed below.

Nguyen (2010) proposed SF models with Cauchy and Laplace distributions for v_i, pairing the former with half Cauchy and truncated Cauchy,

[18]For an explanation of shrinkage in the context of the predictor $E(u_i|\varepsilon_i)$, see Wang and Schmidt (2009).

[19]Holding β constant.

and the latter with exponential and truncated Laplace distributed for has received v_i terms.[20] Gupta and Nguyen (2010) derive a Cauchy-half Cauchy panel data model with time-invariant inefficiency. Horrace and Parmeter (2018) consider the Laplace-truncated Laplace and Laplace-exponential SF models further, showing that $f_{u|\varepsilon}$ (and therefore also $E(u_i|\varepsilon_i)$, or for that matter any predictor based on $f_{u|\varepsilon}$) is constant for $s\varepsilon_i \geq 0$. The authors conjecture that the assumption of a Laplace distributed v_i may be advantageous in terms of estimation of f_u, and therefore for the deconvolution of the composed error. Fan (1991) showed that optimal rates of convergence in deconvolution problems decrease with the smoothness of the noise distribution and are considerably faster for ordinary smooth distributions, such as the Laplace, than for super smooth distributions, such as the normal distribution. Optimal convergence rates for nonparametric Gaussian deconvolution are discussed by Fan (1992). Horrace and Parmeter (2011) find that consistent estimation of the distribution of u_i in a semparametric SF model, in which $v_i \sim N\left(0, \sigma_v^2\right)$ and f_u is unknown, has a $\ln n$ convergence rate. This implies that convergence rates when $v_i \sim N\left(0, \sigma_v^2\right)$ are rather slow.

In the aforementioned proposals, the distribution of u_i is the left truncation at zero of the distribution of v_i. In many cases, this ensures that f_ε can be expressed analytically. Proposition 9 of Azzalini and Capitanio (2003) shows the density of the sum of a random variable and the absolute value of another random variable following the same elliptical distribution. Stead et al. (2018) propose the use of MSL to pair a thick-tailed distribution for v_i with any given distribution for u_i, and estimate a logistic-half normal SF model. The authors show that the model yields a narrower range of efficiency scores compared to the normal-half normal model.

There are two drawbacks of the above proposals for v_i. First, they have fixed shapes, so there is no flexibility in the heaviness of their tails. Second, they do not nest the normal distribution, which makes testing against the standard SF model difficult. One potential noise distribution with neither of these shortcomings is the Student's t distribution, which has a 'degrees of freedom' parameter α that determines the heaviness of the tails, and which approaches the normal distribution as $\alpha \to \infty$. Tancredi (2002) proposed an SF model in which v_i and u_i follow non-standard Student's t distribution and half t distributions, with scale parameters σ_v and σ_u, respectively,

[20] In keeping with previous terminology, 'truncated' (without further qualification) refers specifically to the left truncation at zero of a distribution with mean μ, and 'half' refers to the special case where $\mu = 0$. Note that truncating the Laplace distribution thus yields the exponential distribution whenever $\mu \leq 0$ due to the memorylessness property of the exponential distribution.

and a common degrees of freedom parameter α. The author shows that $f_{u|\varepsilon} \to 0$ as $s\varepsilon_i \to \infty$ and that $E\left[\exp\left(-u_i|\varepsilon_i\right)\right]$ and $\text{VAR}\left[\exp\left(-u_i|\varepsilon_i\right)\right]$ are non-monotonic functions of ε_i. Wheat et al. (2019) estimate a t-half normal model via MSL, similarly finding that $E(u_i|\varepsilon_i)$ is non-monotonic, decreasing with $s\varepsilon_i$ at either tail, and discuss testing against the normal-half normal SF model. Bayesian estimation of the t-half t model, and of t-half normal, t-exponential, and t-gamma SF models are discussed by Tchumtchoua and Dey (2007) and Griffin and Steel (2007), respectively.

Another proposal which nests the standard SF model and allows for flexibility in the kurtosis of v_i, is that of Wheat et al. (2017), in which v_i follows a mixture of two normal distributions with zero means, variances σ^2_{v1} and σ^2_{v2}, respectively, and mixing parameter p. This is often referred to as the contaminated normal distribution.[21] Alternatively, the model can be interpreted as a latent class model with two regimes having differing noise variances. Efficiency prediction in latent class and mixture SF models is discussed, and $E(u_i|\varepsilon_i)$ is shown to be non-monotonic in the contaminated normal-half normal case, as in the t-half normal. Testing down to the standard SF model is less straightforward in this case, since there is an unidentified parameter under the null hypothesis.

The proposals discussed in this section have all been motivated to one degree or another by the need to accommodate outliers in a satisfactory way. An exception to this general rule is Bonanno et al. (2017), who propose an SF model with correlated error components—for a discussion of such models, see Sect. 8.1—in which the marginal distributions of v_i and u_i are skew logistic and exponential, respectively. The motivation in this case is to allow for non-zero efficiency predictions in the presence of 'wrong skew', which the model ascribes to the skewness of v_i.

7 Presence of Efficient Firms

A number of papers have considered SFA in the case where some significant proportion of firms lie on the frontier—i.e. are fully efficient—and discussed SF specifications and efficiency prediction appropriate for this case, along with methods used to identify subset of efficient firms.

Horrace and Schmidt (2000) discuss multiple comparisons with the best (MCB)—see Hsu (1981, 1984) for background on MCB—in which there

[21] Or more specifically, the scale contaminated normal distribution.

are I populations each with their own distinct parameter values, a_i, one of which—e.g. the maximum or the minimum—is the 'best' in some sense, against which we want to compare the remaining I − 1 populations. Rather than make individual comparisons, e.g. by testing $H_0 : a_i = a_b$ where $a_b = \max_{j \neq i} sa_j$), MCB constructs joint confidence intervals for a vector of differences $\left(a_b - a_1 \; a_b - a_2 \; \ldots \; a_b - a_{I-1} \right)$. This is motivated by the need to consider the 'multiplicity effect' (Hochberg and Tamhane 1987), i.e. the fact that if a large enough number of comparisons are made, some differences are bound to appear significant. MCB is also concerned with constructing a set of populations which could be the best. Horrace and Schmidt (2000) discuss application of MCB to derive such multivariate intervals in the context of the fixed effects, time-invariant efficiency panel SF model of Schmidt and Sickles (1984), and the selection of a set of efficient (or probably efficient) firms based on these.

An alternative approach proposed by Jung (2017) is to use a least absolute shrinkage and selection operator (LASSO) variant of the Schmidt and Sickles (1984) model. LASSO is a method used for variable selection and to penalise overfitting by shrinking the parameter estimates towards zero and was introduced by Tibshirani (1996) in the context of OLS, such that

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left[\frac{1}{I} \sum_{i=1}^{I} \varepsilon_i^2 + \lambda \sum_{k=1}^{K} |\beta_k| \right] \tag{25}$$

where K is the number of regressors, and λ is a tuning parameter that determines the strength of the penalty (or the degree of shrinkage). The constant term β_0 is excluded from the penalty term. The penalty is such that it forces some of the coefficients to be zero, hence, its usefulness in variable selection. It is straightforward to extend the approach to a fixed-effects panel data model. Jung (2017) proposes extending the approach to the Schmidt and Sickles (1984) fixed effects SF model, in which $\beta_0 = \max_j sa_j$ and $u_i = \max_j sa_j - sa_i$, and introduces an additional penalty term such that the inefficiency parameters are shrunk towards zero, and $u_i = 0$ for a subset of firms. The author discusses the properties of the model, and in applying the model to a dataset used by Horrace and Schmidt (2000), notes that the resulting set of efficient firms is similar to that obtained using the MCB approach.

Kumbhakar et al. (2013) proposed a zero inefficiency stochastic frontier (ZISF) model. The ZISF model adapts the standard parametric SF model to account for the possibility that a proportion, p, of the firms in the sample

are fully efficient using a latent class approach in which $u_i = 0$ with probability p. That is, the ZISF model is a latent class model in which

$$f_\varepsilon(\varepsilon_i) = pf_v(\varepsilon_i) + (1-p)\int_0^\infty f_v(\varepsilon_i + su_i)f_u(u_i)\mathrm{d}u_i \tag{26}$$

where f_v is the density of v_i and assumed noise distribution, and f_u is the density of u_i in the second regime. In the first regime, u_i can be thought of as belonging to a degenerate distribution at zero. The ZISF model nests the standard SF model when $p = 0$, and testing down to the SF model is a standard problem. On the other hand, testing $H_0 : p = 1$, i.e. that all firms are fully efficient, is more complicated, that the splitting proportion p lies on the boundary of the parameter space in this case. The authors suggest that the LR statistic follows a $\chi^2_{1:0}$ distribution.[22] That is, a 50:50 mixture of χ^2_0 and χ^2_1 distributions. However, Rho and Schmidt (2015) question the applicability of this result, noting an additional complication: under $H_o : p = 1$, σ_u is not identified. Equivalently, p is not identified under $H_o : \sigma_u = 0$. Simulation evidence provided by the authors suggests that estimates of these two parameters are likely to be imprecise when either is small.

Kumbhakar et al. (2013) suggest several approaches to efficiency prediction from the ZISF model. First, the authors suggest weighting regime-specific efficiency predictions by unconditional probabilities of regime membership. Since $\hat{u}_i = 0$ in the first regime regardless of the predictor used, this amounts to using $(1-p)E(u_i|\varepsilon_i)$. This is clearly unsatisfactory, as each firm is assigned the same (unconditional) probabilities for regime membership. A preferable alternative, suggested by both Kumbhakar et al. (2013) and Rho and Schmidt (2015), suggest using $(1-p_i)E(u_i|\varepsilon_i)$, where $p_i = pf_v(\varepsilon_i)/f_\varepsilon(\varepsilon_i)$, which is a firm-specific probability conditional on ε_i. Note that $(1-p)E(u_i|\varepsilon_i)$ and $(1-p_i)E(u_i|\varepsilon_i)$ for all i and any value of ε_i will yield non-zero predictions of u_i under the assumption that $v_i \sim N\left(0, \sigma_v^2\right)$ (see the discussion of the monotonicity properties of $E(u_i|\varepsilon_i)$ in Sect. 6), despite the fact we expect pI efficient firms in the sample. Kumbhakar et al. (2013) suggest identifying firms as efficient when p_i is greater than some cut-off point; however, the choice of such a cut-off point is arbitrary.

[22] As discussed in Sect. 3, see Case 5 in Self and Liang (1987).

Despite the ZISF model's motivation, efficient firms cannot be identified on the basis of the resulting point predictions of efficiency or conditional probabilities of regime membership. Firms may be predicted as fully efficient if the conditional mode predictor is used, or possibly if an alternative distribution for v_i is assumed (again, refer to Sect. 6), but this is equally true in the standard SF context. An appropriate approach to classifying firms would be to identify those with minimum width prediction intervals, analogous to those derived by Wheat et al. (2014) for $u_i|\varepsilon_i$ in the standard SF model, including zero.

There are trade-offs between each of the three proposed methods. Compared to the ZISF model, the MCB and LASSO approaches have the advantage that no particular distribution for u_i is imposed, and efficient firms can be identified on the basis of hypothesis tests. In contrast, the ZISF model limits us to examining prediction intervals. On the other hand, Horrace and Schmidt (2000) and Jung (2017) assume time-invariant efficiency. While Horrace and Schmidt (2000) state that the MCB approach could be adapted to allow for time-varying efficiency (and the same may be true of the LASSO approach), the ZISF approach is the only one that can be applied to cross-sectional data. In addition, it would be straightforward to extend the ZISF approach to incorporate many features found in the SF literature.

8 Miscellaneous Proposals

In this section, we discuss several of the lesser and relatively tangential strands of the SF literature which have adopted novel distributional forms.

8.1 Correlated Errors

A common assumption across all of the aforementioned SF specifications is that the error components, including all noise, inefficiency and random effects components are distributed independently of one another.[23] Relaxing this assumption seems particularly justified in cases in which there are two or more inefficiency components. Independence between noise and inefficiency terms is usually assumed on the basis that noise represents random

[23]Again, as an exception to this, dependency between error components may be introduced via 'environmental' variables influencing the parameters of their distributions as discussed in Sect. 5.

factors unrelated to efficiency. On the other hand, it has been argued that such factors may affect firm decision making and therefore efficiency.

Similar to the panel data case discussed in Sect. 4.1, one approach to modelling dependence between errors has been to specify some multivariate analogue to common distributional assumptions under independence. Schmidt and Lovell (1980), Pal and Sengupta (1999), and Bandyopadhyay and Das (2006) consider a left truncated a bivariate normal distribution at zero with respect to a one-sided inefficiency component.[24] The two-sided component represents noise in the latter two cases and allocative inefficiency in the former. Pal and Sengupta (1999) likewise included allocative inefficiency components, which are assumed to follow a multivariate normal distribution. However, the marginal distributions of the error components are not those commonly used under independence and, more importantly, that they may be inappropriate. Bandyopadhyay and Das (2006) show that while the marginal distribution of u_i in their model is half normal, that of v_i is skew normal, with skewness determined by the correlation between the two error components. An unusual approach was proposed by Pal (2004), in which conditional distributions for the error components are specified directly along with their marginal distributions. Prediction of efficiency is based on $f_{u|\varepsilon}$ as in the case of independence.

The use of a copula function to allow for dependence between v_i and u_i was proposed by Smith (2008) and El Mehdi and Hafner (2014). Various alternatives are considered, including the Ali-Mikhail-Haq, Clayton, Fairlie-Gumbel-Morgenstern, Frank and Gaussian copula. From Sklar's theorem, the joint density $f_{v,u}$ is the product of the marginal densities and the density of the copula. It follows that (3) and (4) must be modified such that

$$f_\varepsilon(\varepsilon_i) = \int_0^\infty f_v(\varepsilon_i + su_i) f_u(u_i) c_{v,u}[F_v(\varepsilon_i + su_i), F_u(u_i)] \mathrm{d}u_i \tag{27}$$

and

$$f_{u_i|\varepsilon_i}(u_i|\varepsilon_i) = \frac{f_v(\varepsilon_i + su_i) f_u(u_i) c_{v,u}[F_v(\varepsilon_i + su_i), F_u(u_i)]}{f_\varepsilon(\varepsilon_i)} \tag{28}$$

[24] Schmidt and Lovell (1980) fold, rather than truncate.

where $c_{v,u}$ is the copula density. Gómez-Déniz and Pérez-Rodríguez (2015) specify a bivariate Sarmanov distribution for v_i and u_i with normal and half-normal marginal distributions, respectively. Again, the advantage of the copula approach is that the desired marginal distributions are obtained, with the dependence between the error components captured by $c_{v,u}$.

8.2 Sample Selection and Endogenous Switching

In the context of linear regression, the sample selection model of Heckman (1976, 1979) is such that

$$y_i = \begin{cases} x_i\beta + \varepsilon_i, & d_i = 1 \\ \text{unobserved}, & d_i = 0 \end{cases}, \quad d_i = I\left(d_i^* = z_i\alpha + w_i > 0\right), \tag{29}$$

where symmetric error terms ε_i and w_i are assumed to follow a bivariate normal distribution with zero means, variances σ_ε^2 and 1, and correlation coefficient ρ. Unless $\rho = 0$, least squares will yield biased estimates. Since $E(y_i|x_i, d_i = 1) = x_i\beta + \rho\sigma_\varepsilon f_w(z_i\alpha)/F_w(z_i\alpha)$, Heckman (1979) proposed a two-step, limited information method in which y_i is regressed on x_i and the inverse Mills' ratio $f_w\left(z_i\hat{\alpha}\right)/F_w\left(z_i\hat{\alpha}\right)$, where $\hat{\alpha}$ is obtained from a single equation probit model estimated by ML. Alternatively, a full information ML approach may be used to estimate the parameters of the model simultaneously, as in Heckman (1976) and Maddala (1983).

A similar problem is that of endogenous switching. The endogenous switching model of Heckman (1978) has two regimes, membership of which is dependent upon a binary switching dummy:

$$y_i = \begin{cases} x_i\beta_1 + \varepsilon_{1i}, & d_i = 1 \\ x_i\beta_2 + \varepsilon_{2i}, & d_i = 0 \end{cases}, \quad d_i = I\left(d_i^* = z_i\alpha + w_i > 0\right) \tag{30}$$

where ε_{1i}, ε_{2i} and w_i are assumed to follow a trivariate normal distribution with zero means, and variances $\sigma_{1\varepsilon}^2$, $\sigma_{2\varepsilon}^2$, and σ_w^2. The correlations of ε_{1i} and ε_{2i} with w_i are given by ρ_1 and ρ_2, respectively, while ρ_{12} is the correlation between ε_{1i} and ε_{2i}. Again, both two-step partial information and full information ML approaches may be used to estimate the parameters of the model.

In recent years, SF models incorporating sample selection and endogenous switching have been proposed. Bradford et al. (2001) and Sipiläinen and Oude Lansink (2005) use the Heckman (1979) two-step

approach, including the estimated inverse Mills' ratios from single equation probit selection and switching models, respectively, as independent variables in their SF models. However, this is inappropriate in non-linear settings such as SFA, since it is generally *not* the case that $E\left[g(x_i\beta+\varepsilon_i)|d_i=1\right]=g\left[x_i\beta+\rho\sigma_\varepsilon f_w(z_i\alpha)/F_w(z_i\alpha)\right]$ where g is some non-linear function. Terza (2009) discusses ML estimation of non-linear models with endogenous switching or sample selection in general.

In the SF context, there are many alternative assumptions that may be made about the relationship between noise, inefficiency, and the stochastic component of the selection (or switching) equation. Perhaps the natural approach, implicit in Bradford et al. (2001) and Sipiläinen and Oude Lansink (2005), is to assume that the symmetric noise terms follow a multivariate normal distribution as in the linear model, while the inefficiency terms are drawn from independent one-sided univariate distributions. This is proposed by Greene (2010), who estimates an SF model with sample selection via MSL, and also by Lai (2015), who uses the result that, in both the sample selection and endogenous switching cases, $f_{\varepsilon|d}$ follows a closed skew-normal distribution when the inefficiency terms are truncated normal. This results in analytical log-likelihoods, and the author proposes to predict efficiency based on the distribution of $u_i|(\varepsilon_i|d_i)$, specifically using $E\left[\exp\ (-u_i)|(\varepsilon_i|d_i)\right]$.

Note that the distributional assumptions in Greene (2010) and Lai (2015) ensure appropriate marginal distributions for each error component, but do not allow for correlation between the inefficiency terms and the symmetric errors. Lai et al. (2009) introduce correlation between ε_i (rather than its components) and w_i through a copula function. Departing from the usual approach, Kumbhakar et al. (2009) propose an SF model with an endogenous switching equation in which $d_i^*=z_i\alpha+\delta u_i+w_i$. That is, they include the inefficiency term as a determinant of regime membership.[25] The various error components are assumed to be independent of one another, and both the log-likelihood of the model and $E[u_i|(\varepsilon_i|d_i)]$ are obtained by quadrature.

[25]Kumbhakar et al. (2009), using panel data, also include a lagged regime membership (i.e. technology choice) dummy in their selection equation.

8.3 Two-Tiered Models

SF methods have been widely applied outside of the context of production and cost frontier estimation. Most applications have utilised standard cross-section or panel data SF specifications, or some of the variants discussed above. However, one area of application which has seen its own distinct methodological developments is modelling of earnings determination. Polachek and Yoon (1987) proposed a 'two-tiered' SF (2TSF) model in which

$$\varepsilon_i = v_i - u_i + w_i, \tag{31}$$

where v_i is again a normally distributed noise component, and u_i and w_i follow exponential distributions with means σ_u and σ_w, respectively.[26] The dependent variable is a worker's actual wage. The u_i component captures deviations from the firm's reservation wage—i.e. the maximum wage offers the firm would make—as a result of incomplete information on the part of the employee. Similarly, w_i captures deviations from the worker's reservation wage—i.e. the minimum wage offer the worker would accept—as a result of incomplete information on the part of the employer. The inclusion of these two terms therefore allows estimation of the extent of average employee and employer incomplete information, and even observation-specific predictions of these. The assumption of exponentially distributed u_i and w_i makes derivation of f_ε, and therefore the log-likelihood, straightforward. However, as in the standard SF model, alternative distributional assumptions have been proposed: Papadopoulos (2015) derive a closed form for f_ε when u_i and w_i follow half-normal distributions, and Tsionas (2012) estimates the model assuming that they follow gamma distributions via inverse fast Fourier transform of the characteristic function as discussed in Sect. 3.

In general, developments of the 2TSF model have tended to parallel those of the standard SF model. A panel data 2TSF model was proposed by Polachek and Yoon (1996), in which

$$\varepsilon_{ift} = v_{ift} - u_{it} + w_{ft} \tag{32}$$

where the subscript f denotes the firm. The employee incomplete information component u_{it} and the employer incomplete information component

[26]The authors actually use u_i to denote the noise term and v_i and w_i for the one-sided errors. In the interest of consistency and to avoid confusion, we use v_i to refer to the noise term and u_i and w_i for the one-sided errors.

w_{ft}, which is assumed to be constant across all employees, are further decomposed such that $u_{it} = u_i + u_{it}^*$ and $w_{ft} = w_f + w_{ft}^*$, where u_i and w_f are time-invariant fixed effects and u_{it}^* and w_{ft}^* follow independent exponential distributions. It is clear that many alternative panel data specifications could be proposed, particularly considering the numerous possible extensions of the models discussed in Sect. 4.

In addition, and analogous to the models discussed in Sect. 5, modelling of u_i and w_i in terms of vectors of explanatory variables has been proposed. Assuming exponential u_i and w_i, Groot and Oosterbeek (1994) propose modelling the inverse signal-to-noise ratios σ_v/σ_u and σ_v/σ_w as linear functions of vectors z_{ui} and z_{wi}. This specification introduces heteroskedasticity of each of the error components, but in rather an odd way, and is problematic in that it does not restrict σ_u or σ_w to be positive. This issue is resolved by Kumbhakar and Parmeter (2010), who propose a specification in which $\sigma_{ui} = \exp(z_{ui}d_u)$ and $\sigma_{wi} = \exp(z_{wi}d_w)$. Note that this model has the scaling property. Parmeter (2018) proposes estimating a 2TSF model with the scaling property, avoiding explicit distributional assumptions, by non-linear least squares.

Finally, tying back to the previous section, Blanco (2017) proposes an extension of the basic Polachek and Yoon (1987) model to account for sample selection, assuming that the symmetric error components follow a bivariate normal distribution, while the one-sided errors follow independent univariate exponential distributions.

9 Conclusion

The methodological literature on SFA has developed considerably since the first SF models were developed by Aigner et al. (1977) and Meeusen and van Den Broeck (1977). The defining feature of SFA models is the focus on determining observation-specific predictions for inefficiency. This in turn requires a prediction of an inefficiency error terms which is present in tandem with a noise error. Hence, there is a deconvolution problem associated with the error in the model. As such, distributional assumptions are not just required to get 'best' estimates of the underlying frontier relationship (cost frontier, production frontier, etc.), but also essential for enabling appropriate predictions of the quantity of interest: firm inefficiency.

This review has considered numerous ways in which SFA has been innovated, which in turn has involved the use of differing distributional forms. One strand of literature concerns alternative distributional assumptions for

the inefficiency error term, and more recently, the noise error term. This raises the obvious question as to which to choose. Given economic theory only requires the inefficiency error to be one-sided, it is generally an empirical matter as to which is to be preferred. Formulations which nest other forms as special cases have obvious appeal; however, there are also non-nested tests, such as those developed by Wang et al. (2011) to aid selection.

Another strand of literature considers alternative distributions in the presence of specific empirical issues. The 'wrong-skew' problem is a good example, where it is entirely plausible that inefficiency could be found to have skewness counter to the direction imposed by the use of the common, half-normal, truncated-normal or exponential inefficiency distributions. Without a change to the distributional assumptions, the model estimation would indicate no evidence of inefficiency which is often difficult to justify in the context of knowledge and other available evidence of the performance of the industries that these techniques are applied to.

Other innovations include models for sample selection, the presence of efficient firms and two-tier SF models. Panel data is a data structure which greatly increases the scope of modelling possibilities. It potentially allows for construction of predictors of inefficiency without appeal to 'full' distributional assumptions on the noise and inefficiency (instead only requiring moment assumptions), by exploiting time persistency in inefficiency. Alternatively, full parametric approaches can be adopted, with the benefit of being able to obtain separate predictions for inefficiency—which may have both time-invariant and time-varying components—and time-invariant unobserved heterogeneity.

Finally, a strand of literature has developed characterising heteroskedasticity in the error components. This is of particular interest as it allows for quantification of the determinants of inefficiency, which is important in beginning to explain why there is a performance gap for a firm in addition to providing a prediction of the size of such a gap. This, in turn can be used by stakeholders to guide implementation of better performance.

Overall it is misleading to think of SFA as representing a single approach to efficiency analysis. Instead, SFA characterises a broad set of models, where different approaches will be relevant given the empirical context. The limited scope of this review has excluded several topics such as nonparametric SF models, Bayesian SF models, metafrontiers, and estimation of distance functions. Inefficiency is an unobserved error component, and so by definition, the predictor of such an error will be sensitive to distributional assumptions regarding inefficiency and the other unobserved error components, such as noise and unobserved heterogeneity. Thus, the conclusion is that for any

given empirical application of efficiency analysis, several SFA models will need to be considered in order to establish the sensitivity of the efficiency predictions to the distributional assumptions adopted. This review should provide a useful starting point for such an exercise.

References

Afriat, S.N. 1972. Efficiency estimation of production functions. *International Economic Review* 13 (3): 568–598.

Ahn, S.C., Y.H. Lee, and P. Schmidt. 2007. Stochastic frontier models with multiple time-varying individual effects. *Journal of Productivity Analysis* 27 (1): 1–12.

Ahn, S.C., Y.H. Lee, and P. Schmidt. 2013. Panel data models with multiple time-varying individual effects. *Journal of Econometrics* 174 (1): 1–14.

Aigner, D.J., and S.F. Chu. 1968. On estimating the industry production function. *The American Economic Review* 58 (4): 826–839.

Aigner, D.J., T. Amemiya, and D.J. Poirier. 1976. On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review* 17 (2): 377–396.

Aigner, D., C.A.K. Lovell, and P. Schmidt. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6 (1): 21–37.

Almanidis, P., and R.C. Sickles. 2012. The skewness issue in stochastic frontiers models: Fact or fiction? In *Exploring research frontiers in contemporary statistics and econometrics: A festschrift for Léopold Simar*, ed. I. Van Keilegom and W.P. Wilson, 201–227. Heidelberg: Physica-Verlag HD.

Almanidis, P., J. Qian, and R.C. Sickles. 2014. Stochastic frontier models with bounded inefficiency. In *Festschrift in Honor of Peter Schmidt: Econometric Methods and Applications*, ed. R.C. Sickles and W.C. Horrace, 47–81. New York, NY: Springer.

Alvarez, A., C. Amsler, L. Orea, and P. Schmidt. 2006. Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics. *Journal of Productivity Analysis* 25 (3): 201–212.

Amemiya, T., and T.E. MaCurdy. 1986. Instrumental-variable estimation of an error-components model. *Econometrica* 54 (4): 869–880.

Amsler, C., A. Prokhorov, and P. Schmidt. 2014. Using copulas to model time dependence in stochastic frontier models. *Econometric Reviews* 33 (5–6): 497–522.

Amsler, C., P. Schmidt, and W.-J. Tsay. 2015. A post-truncation parameterization of truncated normal technical inefficiency. *Journal of Productivity Analysis* 44 (2): 209–220.

Andrews, D.W.K. 1993a. An introduction to econometric applications of empirical process theory for dependent random variables. *Econometric Reviews* 12 (2): 183–216.

Andrews, D.W.K. 1993b. Tests for parameter instability and structural change with unknown change point. *Econometrica* 61 (4): 821–856.

Azzalini, A. 1985. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* 12 (2): 171–178.

Azzalini, A., and A. Capitanio. 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65 (2): 367–389.

Badunenko, O., and S.C. Kumbhakar. 2016. When, where and how to estimate persistent and transient efficiency in stochastic frontier panel data models. *European Journal of Operational Research* 255 (1): 272–287.

Bandyopadhyay, D., and A. Das. 2006. On measures of technical inefficiency and production uncertainty in stochastic frontier production model with correlated error components. *Journal of Productivity Analysis* 26 (2): 165–180.

Barrow, D.F., and A.C. Cohen. 1954. On some functions involving Mill's ratio. *The Annals of Mathematical Statistics* 25 (2): 405–408.

Battese, G.E., and T.J. Coelli. 1988. Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. *Journal of Econometrics* 38 (3): 387–399.

Battese, G.E., and T.J. Coelli. 1992. Frontier production functions, technical efficiency and panel data: With application to paddy farmers in India. *Journal of Productivity Analysis* 3 (1): 153–169.

Battese, G.E., and T.J. Coelli. 1995. A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics* 20 (2): 325–332.

Battese, G.E., and G.S. Corra. 1977. Estimation of a production frontier model: With application to the pastoral zone of Eastern Australia. *Australian Journal of Agricultural Economics* 21 (3): 169–179.

Battese, G.E., A.N. Rambaldi, and G.H. Wan. 1997. A stochastic frontier production function with flexible risk properties. *Journal of Productivity Analysis* 8 (3): 269–280.

Baumol, W.J. 1967. *Business behavior, value and growth,* revised ed. New York: Macmillan.

Beckers, D.E., and C.J. Hammond. 1987. A tractable likelihood function for the normal-gamma stochastic frontier model. *Economics Letters* 24 (1): 33–38.

Belotti, F., and G. Ilardi. 2018. Consistent inference in fixed-effects stochastic frontier models. *Journal of Econometrics* 202 (2): 161–177.

Bera, A.K., and S.C. Sharma. 1999. Estimating production uncertainty in stochastic frontier production function models. *Journal of Productivity Analysis* 12 (3): 187–210.

Blanco, G. 2017. Who benefits from job placement services? A two-sided analysis. *Journal of Productivity Analysis* 47 (1): 33–47.

Bonanno, G., D. De Giovanni, and F. Domma. 2017. The 'wrong skewness' problem: A re-specification of stochastic frontiers. *Journal of Productivity Analysis* 47 (1): 49–64.

Bracewell, R.N. 1978. *The Fourier transform and its applications*, 2nd ed. New York: McGraw-Hill.

Bradford, W.D., A.N. Kleit, M.A. Krousel-Wood, and R.N. Re. 2001. Stochastic frontier estimation of cost models within the hospital. *The Review of Economics and Statistics* 83 (2): 302–309.

Brorsen, B.W., and T. Kim. 2013. Data aggregation in stochastic frontier models: the closed skew normal distribution. *Journal of Productivity Analysis* 39 (1): 27–34.

Carree, M.A. 2002. Technological inefficiency and the skewness of the error component in stochastic frontier analysis. *Economics Letters* 77 (1): 101–107.

Caudill, S.B., and J.M. Ford. 1993. Biases in frontier estimation due to heteroscedasticity. *Economics Letters* 41 (1): 17–20.

Caudill, S.B., J.M. Ford, and D.M. Gropper. 1995. Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity. *Journal of Business & Economic Statistics* 13 (1): 105–111.

Chamberlain, G. 1984. Panel data. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, 1247–1318. Amsterdam: Elsevier.

Chen, Y.-Y., P. Schmidt, and H.-J. Wang. 2014. Consistent estimation of the fixed effects stochastic frontier model. *Journal of Econometrics* 181 (2): 65–76.

Coelli, T. 1995. Estimators and hypothesis tests for a stochastic frontier function: A Monte Carlo analysis. *Journal of Productivity Analysis* 6 (3): 247–268.

Coelli, T.J., D.S.P. Rao, and G.E. Battese. 2005. *An introduction to efficiency and productivity analysis*, 2nd ed. New York: Springer.

Colombi, R., S.C. Kumbhakar, G. Martini, and G. Vittadini. 2014. Closed-skew normality in stochastic frontiers with individual effects and long/short-run efficiency. *Journal of Productivity Analysis* 42 (2): 123–136.

Cornwell, C., P. Schmidt, and R.C. Sickles. 1990. Production frontiers with cross-sectional and time-series variation in efficiency levels. *Journal of Econometrics* 46 (1): 185–200.

Cuesta, R.A. 2000. A production model with firm-specific temporal variation in technical inefficiency: With application to Spanish dairy farms. *Journal of Productivity Analysis* 13 (2): 139–158.

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

Deprins, D., and L. Simar. 1989a. Estimating technical inefficiencies with correction for environmental conditions. *Annals of Public and Cooperative Economics* 60 (1): 81–102.

Deprins, D., and L. Simar. 1989b. Estimation de frontières déterministes avec facteurs exogènes d'inefficacitè. *Annales d'Économie et de Statistique* 14: 117–150.

El Mehdi, R., and C.M. Hafner. 2014. Inference in stochastic frontier analysis with dependent error terms. *Mathematics and Computers in Simulation* 102 (Suppl. C): 104–116.

Fan, J. 1991. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics* 19 (3): 1257–1272.

Fan, J. 1992. Deconvolution with supersmooth distributions. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* 20 (2): 155–169.

Farrell, M.J. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)* 120 (3): 253–290.

Filippini, M., and W. Greene. 2016. Persistent and transient productive inefficiency: a maximum simulated likelihood approach. *Journal of Productivity Analysis* 45 (2): 187–196.

Gómez-Déniz, E., and J.V. Pérez-Rodríguez. 2015. Closed-form solution for a bivariate distribution in stochastic frontier models with dependent errors. *Journal of Productivity Analysis* 43 (2): 215–223.

González-Farías, G., J.A. Domínguez-Molina, and A.K. Gupta. 2004a. Additive properties of skew normal random vectors. *Journal of Statistical Planning and Inference* 126 (2): 521–534.

González-Farías, G., J.A. Domínguez-Molina, and A.K. Gupta. 2004b. The closed skew-normal distribution. In *Skew-elliptical distributions and their applications*, ed. M.G. Genton, 25–42. Boca Raton: Chapman and Hall/CRC.

Gouriéroux, C., A. Holly, and A. Monfort. 1982. Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica* 50 (1): 63–80.

Greene, W.H. 1980. Maximum likelihood estimation of econometric frontier functions. *Journal of Econometrics* 13 (1): 27–56.

Greene, W.H. 1990. A gamma-distributed stochastic frontier model. *Journal of Econometrics* 46 (1): 141–163.

Greene, W.H. 2003. Simulated likelihood estimation of the normal-gamma stochastic frontier function. *Journal of Productivity Analysis* 19 (2): 179–190.

Greene, W.H. 2004. Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Economics* 13 (10): 959–980.

Greene, W.H. 2005a. Fixed and random effects in stochastic frontier models. *Journal of Productivity Analysis* 23 (1): 7–32.

Greene, W.H. 2005b. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics* 126 (2): 269–303.

Greene, W.H. 2008. The econometric approach to efficiency analysis. In *The measurement of productive efficiency and productivity growth*, 2nd ed., ed. H.O. Fried, C.A.K. Lovell, and S.S. Schmidt, 92–159. Oxford: Oxford University Press.

Greene, W.H. 2010. A stochastic frontier model with correction for sample selection. *Journal of Productivity Analysis* 34 (1): 15–24.

Greene, W.H. 2016. *LIMDEP Version 11.0 econometric modeling guide.* Econometric Software.

Greene, W.H., and S. Misra. 2003. Simulated maximum likelihood estimation of general stochastic frontier regressions. Working Paper, William Simon School of Business, University of Rochester.

Griffin, J.E., and M.F.J. Steel. 2007. Bayesian stochastic frontier analysis using WinBUGS. *Journal of Productivity Analysis* 27 (3): 163–176.

Griffin, J.E., and M.F.J. Steel. 2008. Flexible mixture modelling of stochastic frontiers. *Journal of Productivity Analysis* 29 (1): 33–50.

Griffiths, W.E., and G. Hajargasht. 2016. Some models for stochastic frontiers with endogeneity. *Journal of Econometrics.* 190 (2): 341–348.

Groot, W., and H. Oosterbeek. 1994. Stochastic reservation and offer wages. *Labour Economics* 1 (3): 383–390.

Grushka, E. 1972. Characterization of exponentially modified Gaussian peaks in chromatography. *Analytical Chemistry* 44 (11): 1733–1738.

Gupta, A.K., and N. Nguyen. 2010. Stochastic frontier analysis with fat-tailed error models applied to WHO health data. *International Journal of Innovative Management, Information & Production* 1 (1): 43–48.

Hadri, K. 1999. Estimation of a doubly heteroscedastic stochastic frontier cost function. *Journal of Business & Economic Statistics* 17 (3): 359–363.

Hajargasht, G. 2014. The folded normal stochastic frontier model. Working Paper.

Hajargasht, G. 2015. Stochastic frontiers with a Rayleigh distribution. *Journal of Productivity Analysis* 44 (2): 199–208.

Hansen, B.E. 1996. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64 (2): 413–430.

Hausman, J.A. 1978. Specification tests in econometrics. *Econometrica* 46 (6): 1251–1271.

Hausman, J.A., and W.E. Taylor. 1981. A generalized specification test. *Economics Letters* 8 (3): 239–245.

Heckman, J.J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5 (4): 17.

Heckman, J.J. 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46 (4): 931–959.

Heckman, J.J. 1979. Sample selection bias as a specification error. *Econometrica* 47 (1): 153–161.

Heshmati, A., and S.C. Kumbhakar. 1994. Farm heterogeneity and technical efficiency: Some results from Swedish dairy farms. *Journal of Productivity Analysis* 5 (1): 45–61.

Hicks, J.R. 1935. Annual survey of economic theory: The theory of monopoly. *Econometrica* 3 (1): 1–20.

Hochberg, Y., and A.C. Tamhane. 1987. *Multiple comparison procedures*, 1st ed. New York: Wiley.

Horrace, W.C. 2005. Some results on the multivariate truncated normal distribution. *Journal of Multivariate Analysis* 94 (1): 209–221.

Horrace, W.C. 2015. Moments of the truncated normal distribution. *Journal of Productivity Analysis* 43 (2): 133–138.

Horrace, W.C., and C.F. Parmeter. 2011. Semiparametric deconvolution with unknown error variance. *Journal of Productivity Analysis* 35 (2): 129–141.

Horrace, W.C., and C.F. Parmeter. 2018. A Laplace stochastic frontier model. *Econometric Reviews* 37 (3): 260–280.

Horrace, W.C., and P. Schmidt. 2000. Multiple comparisons with the best, with economic applications. *Journal of Applied Econometrics* 15 (1): 1–26.

Hsu, J.C. 1981. Simultaneous confidence intervals for all distances from the "best". *The Annals of Statistics* 9 (5): 1026–1034.

Hsu, J.C. 1984. Constrained simultaneous confidence intervals for multiple comparisons with the best. *The Annals of Statistics* 12 (3): 1136–1144.

Huang, C.J., and J.-T. Liu. 1994. Estimation of a non-neutral stochastic frontier production function. *Journal of Productivity Analysis* 5 (2): 171–180.

Johnston, J. 1960. *Statistical cost analysis*, 1st ed. New York: McGraw-Hill.

Jondrow, J., C.A. Knox Lovell, I.S. Materov, and P. Schmidt. 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics* 19 (2): 233–238.

Jovanovic, B., and Y. Nyarko. 1996. Learning by doing and the choice of technology. *Econometrica* 64 (6): 1299–1310.

Jung, H. 2017. Adaptive LASSO for stochastic frontier models with many efficient firms. Working Paper, Maxwell School of Citizenship and Public Affairs, Syracuse University.

Just, R.E., and R.D. Pope. 1978. Stochastic specification of production functions and economic implications. *Journal of Econometrics* 7 (1): 67–86.

Kneip, A., R.C. Sickles, and W. Song. 2012. A new panel data treatment for heterogeneity in time trends. *Econometric Theory* 28 (3): 590–628.

Kodde, D.A., and F.C. Palm. 1986. Wald criteria for jointly testing equality and inequality restrictions. *Econometrica* 54 (5): 1243–1248.

Koopmans, T.C. 1951. Efficient allocation of resources. *Econometrica* 19 (4): 455–465.

Kumbhakar, S.C. 1990. Production frontiers, panel data, and time-varying technical inefficiency. *Journal of Econometrics* 46 (1): 201–211.

Kumbhakar, S.C. 1991. Estimation of technical inefficiency in panel data models with firm- and time-specific effects. *Economics Letters* 36 (1): 43–48.

Kumbhakar, S.C. 1993. Production risk, technical efficiency, and panel data. *Economics Letters* 41 (1): 11–16.

Kumbhakar, S.C., and A. Heshmati. 1995. Efficiency measurement in Swedish dairy farms: An application of rotating panel data, 1976–88. *American Journal of Agricultural Economics* 77 (3): 660–674.

Kumbhakar, S.C., and L. Hjalmarsson. 1995. Labour-use efficiency in Swedish social insurance offices. *Journal of Applied Econometrics* 10 (1): 33–47.

Kumbhakar, S.C., and C.A.K Lovell. 2000. *Stochastic frontier analysis*, 1st ed. Cambridge: Cambridge University Press.

Kumbhakar, S.C., and C.F. Parmeter. 2010. Estimation of hedonic price functions with incomplete information. *Empirical Economics* 39 (1): 1–25.

Kumbhakar, S.C., and K. Sun. 2013. Derivation of marginal effects of determinants of technical inefficiency. *Economics Letters* 120 (2): 249–253.

Kumbhakar, S.C., S. Ghosh, and J.T. McGuckin. 1991. A generalized production frontier approach for estimating determinants of inefficiency in U.S. dairy farms. *Journal of Business & Economic Statistics* 9 (3): 279–286.

Kumbhakar, S.C., E.G. Tsionas, and T. Sipiläinen. 2009. Joint estimation of technology choice and technical efficiency: an application to organic and conventional dairy farming. *Journal of Productivity Analysis* 31 (3): 151–161.

Kumbhakar, S.C., C.F. Parmeter, and E.G. Tsionas. 2013. A zero inefficiency stochastic frontier model. *Journal of Econometrics* 172 (1): 66–76.

Kumbhakar, S.C., G. Lien, and J.B. Hardaker. 2014. Technical efficiency in competing panel data models: A study of Norwegian grain farming. *Journal of Productivity Analysis* 41 (2): 321–337.

Lai, H.-P. 2015. Maximum likelihood estimation of the stochastic frontier model with endogenous switching or sample selection. *Journal of Productivity Analysis* 43 (1): 105–117.

Lai, H.-P., and C.J. Huang. 2013. Maximum likelihood estimation of seemingly unrelated stochastic frontier regressions. *Journal of Productivity Analysis* 40 (1): 1–14.

Lai, H.-P., S.W. Polachek, and H.-J. Wang. 2009. Estimation of a stochastic frontier model with a sample selection problem. Working Paper, Department of Economics, National Chung Cheng University.

Lee, L.-F. 1983. A test for distributional assumptions for the stochastic frontier functions. *Journal of Econometrics* 22 (3): 245–267.

Lee, L.-F. 1993. Asymptotic distribution of the maximum likelihood estimator for a stochastic frontier function model with a singular information matrix. *Econometric Theory* 9 (3): 413–430.

Lee, Y.H. 1996. Tail truncated stochastic frontier models. *Journal of Economic Theory and Econometrics* 2: 137–152.

Lee, L.-F., and A. Chesher. 1986. Specification testing when score test statistics are identically zero. *Journal of Econometrics* 31 (2): 121–149.

Lee, S., and Y.H. Lee. 2014. Stochastic frontier models with threshold efficiency. *Journal of Productivity Analysis* 42 (1): 45–54.

Lee, Y.H., and P. Schmidt. 1993. A production frontier model with flexible temporal variation in technical efficiency. In *The measurement of productive efficiency: Techniques and applications*, ed. H.O. Fried, S.S. Schmidt, and C.A.K. Lovell, 237–255. Oxford: Oxford University Press.

Lee, L.-F., and W.G. Tyler. 1978. The stochastic frontier production function and average efficiency. *Journal of Econometrics* 7 (3): 385–389.

Leibenstein, H. 1966. Allocative efficiency vs. "X-efficiency". *The American Economic Review* 56(3), 392–415.

Leibenstein, H. 1975. Aspects of the X-efficiency theory of the firm. *The Bell Journal of Economics* 6 (2): 580–606.

Leibenstein, H. 1978. X-inefficiency Xists: Reply to an Xorcist. *The American Economic Review* 68 (1): 203–211.

Li, Q. 1996. Estimating a stochastic production frontier when the adjusted error is symmetric. *Economics Letters* 52 (3): 221–228.

Lukacs, E. and R.G. Laha. 1964. *Applications of characteristic functions.* London: Charles Griffin and Company.

Maddala, G.S. 1983. *Limited-dependent and qualitative variables in econometrics.* Cambridge: Cambridge University Press.

Marris, R.L. 1964. *The economic theory of managerial capitalism*, 1st ed. London: Macmillan.

Meeusen, W., and J. van Den Broeck. 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* 18 (2): 435–444.

Migon, H.S., and E.V. Medici. 2001. Bayesian hierarchical models for stochastic production frontier. Working Paper, Universidade Federal do Rio de Janeiro.

Mundlak, Y. 1978. On the pooling of time series and cross section data. *Econometrica* 46 (1): 69–85.

Murillo-Zamorano, L.R. 2004. Economic efficiency and frontier techniques. *Journal of Economic Surveys* 18 (1): 33–77.

Nelsen, R.B. 2006. *An introduction to copulas*, 2nd ed. New York: Springer.

Nerlove, M. 1963. Returns to scale in electricity supply. In *Measurement in economics: Studies in mathematical economics and econometrics*, ed. C.F. Christ, M. Friedman, L.A. Goodman, Z. Griliches, A.C. Harberger, N. Liviatan, J. Mincer, Y. Mundlak, M. Nerlove, D. Patinkin, L.G. Telser, and H. Theil, 167–198. Stanford: Stanford University Press.

Nguyen, N. 2010. Estimation of technical efficiency in stochastic frontier analysis. PhD thesis, Bowling Green State University.

Oikawa, K. 2016. A microfoundation for stochastic frontier analysis. *Economics Letters* 139: 15–17.

Ondrich, J., and J. Ruggiero. 2001. Efficiency measurement in the stochastic frontier model. *European Journal of Operational Research* 129 (2): 434–442.

Orea, L., and S.C. Kumbhakar. 2004. Efficiency measurement using a latent class stochastic frontier model. *Empirical Economics* 29 (1): 169–183.

Pal, M. 2004. A note on a unified approach to the frontier production function models with correlated non-normal error components: The case of cross section data. *Indian Economic Review* 39 (1): 7–18.

Pal, M., and A. Sengupta. 1999. A model of FPF with correlated error components: An application to Indian agriculture. *Sankhyā: The Indian Journal of Statistics, Series B* 61 (2): 337–350.

Papadopoulos, A. 2015. The half-normal specification for the two-tier stochastic frontier model. *Journal of Productivity Analysis* 43 (2): 225–230.

Parmeter, C.F. 2018. Estimation of the two-tiered stochastic frontier model with the scaling property. *Journal of Productivity Analysis* 49 (1): 37–47.

Parmeter, C.F., and S.C. Kumbhakar. 2014. Efficiency analysis: A primer on recent advances. *Foundations and Trends in Econometrics* 7 (3–4): 191–385.

Pitt, M.M., and L.-F. Lee. 1981. The measurement and sources of technical inefficiency in the Indonesian weaving industry. *Journal of Development Economics* 9 (1): 43–64.

Polachek, S.W., and B.J. Yoon. 1987. A two-tiered earnings frontier estimation of employer and employee information in the labor market. *The Review of Economics and Statistics* 69 (2): 296–302.

Polachek, S.W., and B.J. Yoon. 1996. Panel estimates of a two-tiered earnings frontier. *Journal of Applied Econometrics* 11 (2): 169–178.

Qian, J., and R.C. Sickles. 2008. Stochastic frontiers with bounded inefficiency. Mimeo. Department of Economics, Rice University.

Reifschneider, D., and R. Stevenson. 1991. Systematic departures from the frontier: A framework for the analysis of firm inefficiency. *International Economic Review* 32 (3): 715–723.

Rho, S., and P. Schmidt. 2015. Are all firms inefficient? *Journal of Productivity Analysis* 43 (3): 327–349.

Richmond, J. 1974. Estimating the efficiency of production. *International Economic Review* 15 (2): 515–521.

Ritter, C., and L. Simar. 1997. Pitfalls of normal-gamma stochastic frontier models. *Journal of Productivity Analysis* 8 (2): 167–182.

Schmidt, P. 1976. On the statistical estimation of parametric frontier production functions. *The Review of Economics and Statistics* 58 (2): 238–239.

Schmidt, P., and C.A.K. Lovell. 1980. Estimating stochastic production and cost frontiers when technical and allocative inefficiency are correlated. *Journal of Econometrics* 13 (1): 83–100.

Schmidt, P., and R.C. Sickles. 1984. Production frontiers and panel data. *Journal of Business & Economic Statistics.* 2 (4): 367–374.

Self, S.G., and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82 (398): 605–610.

Shephard, R.W. 1953. *Cost and production functions*, 1st ed. Princeton: Princeton University Press.

Simar, L., and P.W. Wilson. 2010. Inferences from cross-sectional, stochastic frontier models. *Econometric Reviews* 29 (1): 62–98.

Simar, L., C.A.K. Lovell, and P. Vanden Eeckaut. 1994. Stochastic frontiers incorporating exogenous influences on efficiency. STAT Discussion Papers no. 9403, Institut de Statistique, Université Catholique de Louvain.

Sipiläinen, T., and A. Oude Lansink. 2005. Learning in switching to organic farming. *NJF Report* 1 (1): 169–172.

Smith, M.D. 2008. Stochastic frontier models with dependent error components. *Econometrics Journal* 11 (1): 172–192.

Smith, A.S.J., and P. Wheat. 2012. Estimation of cost inefficiency in panel data models with firm specific and sub-company specific effects. *Journal of Productivity Analysis* 37 (1): 27–40.

Stead, A.D. 2017. Regulation and efficiency in UK public utilities. PhD thesis, University of Hull.

Stead, A.D., P. Wheat, and W.H. Greene. 2018. Estimating efficiency in the presence of extreme outliers: A logistic-half normal stochastic frontier model with application to highway maintenance costs in England. In *Productivity and inequality*, ed W.H. Greene, L. Khalaf, P. Makdissi, R.C. Sickles, M. Veall, and M. Voia, 1–19. Springer.

Stevenson, R.E. 1980. Likelihood functions for generalized stochastic frontier estimation. *Journal of Econometrics* 13 (1): 57–66.

Stigler, G.J. 1976. The Xistence of X-efficiency. *The American Economic Review* 66 (1): 213–216.

Tancredi, A. 2002. Accounting for heavy tails in stochastic frontier models. Working Paper no. 2002.16, Department of Statistical Sciences, University of Padua.

Tchumtchoua, S., and D.K. Dey. 2007. Bayesian estimation of stochastic frontier models with multivariate skew *t* error terms. *Communications in Statistics - Theory and Methods* 36 (5): 907–916.

Terza, J.V. 2009. Parametric nonlinear regression with endogenous switching. *Econometric Reviews* 28 (6): 555–580.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–288.

Timmer, C.P. 1971. Using a probabilistic frontier production function to measure technical efficiency. *Journal of Political Economy* 79 (4): 776–794.

Train, K.E. 2009. Discrete choice methods with simulation, 2nd ed. Cambridge: Cambridge University Press.

Tsagris, M., C. Beneki, and H. Hassani. 2014. On the folded normal distribution. *Mathematics* 2 (1): 12–28.

Tsionas, E.G. 2002. Stochastic frontier models with random coefficients. *Journal of Applied Econometrics* 17 (2): 127–147.

Tsionas, E.G. 2007. Efficiency measurement with the Weibull stochastic frontier. *Oxford Bulletin of Economics and Statistics* 69 (5): 693–706.

Tsionas, E.G. 2012. Maximum likelihood estimation of stochastic frontier models by the Fourier transform. *Journal of Econometrics* 170 (1): 234–248.

Tsionas, M.G. 2017. Microfoundations for stochastic frontiers. *European Journal of Operational Research* 258 (3): 1165–1170.

Tsionas, E.G., and S.C. Kumbhakar. 2014. Firm heterogeneity, persistent and transient technical inefficiency: a generalized true random-effects model. *Journal of Applied Econometrics* 29 (1): 110–132.

Waldman, D.M. 1984. Properties of technical efficiency estimators in the stochastic frontier model. *Journal of Econometrics* 25 (3): 353–364.

Wang, H.-J. 2002. Heteroscedasticity and Non-monotonic efficiency effects of a stochastic frontier model. *Journal of Productivity Analysis* 18 (3): 241–253.

Wang, J., and P. Schmidt. 2002. One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis* 18: 129–144.

Wang, W.S., and P. Schmidt. 2009. On the distribution of estimated technical efficiency in stochastic frontier models. *Journal of Econometrics* 148 (1): 36–45.

Wang, W.S., C. Amsler, and P. Schmidt. 2011. Goodness of fit tests in stochastic frontier models. *Journal of Productivity Analysis* 35 (2): 95–118.

Weinstein, M.A. 1964. The sum of values from a normal and a truncated normal distribution. *Technometrics* 6 (1): 104–105.

Wheat, P., and A. Smith. 2012. Is the choice of (t−T) in Battese and Coelli (1992) type stochastic frontier models innocuous? Observations and generalisations. *Economics Letters* 116 (3): 291–294.

Wheat, P., W. Greene, and A. Smith. 2014. Understanding prediction intervals for firm specific inefficiency scores from parametric stochastic frontier models. *Journal of Productivity Analysis* 42 (1): 55–65.

Wheat, P., A.D. Stead, and W.H. Greene. 2017. Allowing for outliers in stochastic frontier models: A mixture noise distribution approach. 15th European Workshop on Efficiency and Productivity Analysis, London, UK.

Wheat, P., A.D. Stead, and W.H. Greene. 2019. Robust stochastic frontier analysis: A student's t-half normal model with application to highway maintenance costs in England. *Journal of Productivity Analysis* 51 (1): 21–38. https://doi.org/10.1007/s11123-018-0541-y.

Williamson, O.E. 1963. Managerial discretion and business behavior. *The American Economic Review* 53 (5): 1032–1057.

Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* 57 (298): 348–368.

Stochastic Frontier Models for Discrete Output Variables

Eduardo Fé

1 Introduction

Stochastic frontier models (SFM) (Aigner et al. 1977; Meeusen and van den Broeck 1977) are central to the identification of inefficiencies in the production of continuously distributed outputs. There are, however, numerous situations where the outcome variable of interest is discretely distributed. This is particularly true in the realm of labour, industrial and health economics where the concept of production frontier has been adopted to explain non-tangible and non-pecuniary outcomes which are often measured through indicators of achievement (employment status, academic certification, success in a labour market scheme), ordered categories (Likert scales describing job satisfaction, health status, personality traits) and counts (the number of patents obtained by a firm or the number of infant deaths in a region). Although these latter fields of inquiry have not emphasised the idea of inefficiency in the production of non-tangible and non-pecuniary outcomes, recent contributions (e.g. Fe 2013; Fe and Hofler 2013) suggest that inefficiencies are also present in these domains. The question is then whether traditional continuous data methods are still suitable in this setting.

There are two main reasons why new models of stochastic frontiers might be required. First, it is well known that imposing parametric restrictions

E. Fé (✉)
Department of Social Statistics, University of Manchester, Manchester, UK
e-mail: Eduardo.FE@manchester.ac.uk

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_9

which take into account the specific features of distributions results in more efficient estimates and better inference (e.g. Greene 2004b). Second, and more importantly, inefficiency is typically not nonparametrically identified from data alone. Therefore, researchers have to make specific assumptions regarding the distribution of inefficiency in the sample or the population. These assumptions define the class of admissible distribution underlying outputs. Standard continuous data models attribute any negative (positive) skewness in the sample to inefficiencies in the production of economic goods (bads). The distributions of discrete outcomes are, however, typically skewed even in the absence of inefficiencies and the sign of skewness is generally independent of whether one is studying an economic good or an economic bad. The consequences of this are exemplified in Fé and Hofler (2013). These authors describe how standard SFM can fail to detect any inefficiency in production when the outcome of interest is a count—even when the underlying inefficiency might be substantial.

In this chapter, we provide a survey of recent contributions to the area of SFM for the analysis of discrete outcomes and we discuss models for count data, binary outcomes and ordered categorical outcomes—as well as a few extensions of these models. Existing contributions to this area are described in Sect. 2 of this chapter. In Sect. 3, we provide a general framework encompassing existing continuous and discrete outcome frontier models. We note that existing models are mixtures of a standard distribution (such as the normal or the Poisson distributions) with an asymmetric distribution (such as the half-normal or gamma distributions), and we take advantage of this feature to provide a general encompassing framework for these models. Specific choices of distributions result in different models, only a few of which have a closed form. Therefore, drawing from Greene (2003), we suggest the use of maximum simulated likelihoods to estimate the parameters of each model. Section 3 further describes the estimation of cross-sectional inefficiency scores using the methodology in Jondrow et al. (1982) and specific implementations of the framework. We consider the Count Data Stochastic Frontier (CDSF) model of Fé and Hofler (2013), a generalisation of Ghadge's logit SFM (Ghadge 2017) and a frequentist version of the ordered logit SFM in Griffiths et al. (2014). For the latter two models, we provide a due account of the conditions necessary for identification. Section 4 of the chapter considers extensions of the general framework. We first discuss the conditions under which unobserved heterogeneity can be introduced in the model. This is an important extension which has been considered (for the count data stochastic model) in Drivas et al. (2014). We show

through simulation which are the consequences of ignoring unobserved heterogeneity in discrete outcome SFM. The extension is also important insofar as it constitutes a prior step in the introduction of endogenous variables in the models—a extension which is, nonetheless, beyond the scope of this chapter. As in Fé (2013) and Fé and Hofler (2013), Sect. 4 further considers semi-nonparametric estimation via local likelihood methods (Tibshirani and Hastie 1987). This extension allows researches to relax the various distributional assumptions imposed in their models, thus reducing the risk of misspecification. Finally, Sect. 5 concludes with some remarks.

2 Stochastic Frontier Models When Outcomes Are Discrete-Valued

The production function (or production frontier) defines the average upper bound for the amount of output achievable with a given set of inputs. More broadly, it summarises the constraints that nature imposes on firms' ability to produce (Wicksteed 1894). Understanding the nature and magnitude of these constraints can result in better planning and targeted interventions to improve production. Therefore, the estimation of production functions is a topic of considerable importance for firms, researchers and policy-makers alike.

Early empirical studies (e.g. Cobb and Douglas 1928) assumed that firms operated on the frontier so that observed variation in output at any level of input was due to a zero-mean random process. This interpretation facilitated the estimation of the production function through conventional methods such as least squares. However, economists soon pointed out that much of the variation in output across firms was likely to be the result of structural factors which lead to production levels below the frontier, such as environmental incentives (derived, for example, from monopolistic advantages—Hicks 1935; Chamberlin 1933) and technical factors (Debreu 1959; Koopmans 1951). To accommodate these inefficiencies, Debreu (1959) and Shephard (1970) introduced the concept of distance function, which extends the scope of production frontiers by decomposing variation in output into the continuously distributed zero-mean error implicit in Wicksteed's original definition (Wicksteed 1894) and a one-sided, continuously distributed error term shifting production away from the frontier.[1]

[1]In what follows, we will take the liberty of abusing terminology and use the terms production function, production frontier and distance function interchangeably.

This development constituted the key stepping stone for the development of the SFM in Aigner et al. (1977) and Meeusen and van den Broeck (1977), which are based on the allocation of probability distributions to each of the error components of a distance function (and are described in a different chapter in this volume).

In parallel to the development of the stochastic frontier literature, labour and industrial economists begun to adopt the idea of a production frontier to explain observed levels of non-tangible, but economically relevant outputs such as human capital (Ben-Porath 1967; Heckman 1976; Mincer 1981), intellectual property (Hall et al. 1986; Wang et al. 1998) and health (Grossman 1972). Empirical research in these areas has not emphasised the idea of technical or environmental inefficiencies,[2] although data often reveal substantial variability in outcomes conditional on the levels of a set of explanatory variables (with much of that variation being attributable to inefficiencies). For example, Fé (2013) estimates that of all the infant deaths reported by any English local authority in 2006, on average 2.5 were due to inefficiencies not accounted for by education, air pollution, smoking behaviour, education or broader socio-economic status.

Applying insights from the production and stochastic frontier literatures to the study of inefficiencies in this type of unconventional setting is, in principle, straightforward. However, an important feature in the literature about the production of intangible outputs is the prevalence of discretely distributed outcomes. For example, intellectual property is often measured through the number of patents awarded to a firm (e.g. Hottenrott et al. 2016) and health is often measured through self-reported scores in ordered Likert scales (e.g. Wildman and Jones 2008; Greene and Hensher 2010). This feature raises a question regarding the extent to which existing stochastic frontier methods can be used to study inefficiencies in the production of discrete outcomes.

2.1 Skewness in the Distribution of Output

Conditional on the level of inefficiency, estimation of mean production frontiers is subject to a trade-off between robustness and statistical efficiency. Estimation of the conditional frontier can be done without employing distributional assumptions using, for example, the ordinary

[2]For a recent example see, e.g., Attanasio et al. (2015).

least squares estimator, which is consistent regardless of the distribution of output. This would justify, for instance, estimation of a mean regression function by least squares with a binary or discrete dependent variable. Assumptions about the actual distribution of output can, however, lead to more efficient estimates and more powerful inferential procedures, but at the expense of increased risk of model misspecification (e.g. White 1982; Greene 2004b; Angrist and Pischke 2008). The question is whether a similar trade-off is available, unconditionally, when fitting continuous data SFM to a discrete outcome.

Inefficiency is not nonparametrically identified from data alone and, as a result, it is essential to draw assumptions about the distribution of inefficiency in the population under study. These assumptions are binding insofar as they define the class of feasible models that can explain variation in output. In the methodology developed by Aigner et al. (1977), there is a clear association between the nature of the output under consideration (an economic good or an economic bad) and the skewness of its distribution (measured by the third central moment, $\kappa_3 = E(Y - E(Y))^3$). In the presence of inefficiency, the distribution of economic goods should exhibit negative skewness, whereas the distribution of economic bads should exhibit positive skewness. Discrete data, however, tend to violate these distributional assumptions. Consider, for example, a discrete output known to have a Poisson distribution. The third moment of this distribution is $\kappa_3 = E(y|x)^{1/2} > 0$. Because the distribution of y is always positively skewed, the assumptions underlying ALS methodology will be violated when y is an economic good. The implication of this violation is that the maximum likelihood estimator of the ALS model will converge to the ordinary least squares estimator, thus suggesting zero inefficiency—even though inefficiency in the sample might be substantial (e.g. Fé and Hoer 2013). As another example, consider the Bernoulli distribution (which serves as the building block for the probit and logit models). Skewness in this case is given by the expression $\kappa_3 = (1 - 2p)/\sqrt{p(1-p)}$ where $p = \text{Prob}(y = 1)$. In this case, skewness does not depend on the nature of the underlying outcome, but rather on the probability that $y = 1$ so it is possible to observe economic goods (or bads) whose distribution is positively or negatively skewed.

The implication of this discussion is that, unlike in non-linear regression, the robustness-efficiency trade-off does not seem to be available. Therefore, specific models are required in order to extend the stochastic frontier methodology to discrete outcomes.

2.2 Existing Contributions

To the best of our knowledge, the literature on SFM for discrete outcomes started with the CDSF model in Fé (2007, 2013). These studies focused on the measurement of regional inequality in the distribution of infant deaths in 350 English local areas. More specifically, the count of infant deaths in a region, y, was treated as the output in an additive distance function, $y = y^* + u$, where y^* is a discrete-valued stochastic frontier and u is a discrete inefficiency term independently distributed from y^*. The distribution of y in this model is given by the convolution of the distributions of y^* and u,

$$P_{y|x}(y_i|x_i;\theta) = \sum_{y^*=0}^{y_i} P_{y^*|x_i}\left(y_i^*|x_{1,i};\theta_1\right) P_{u_i|x_i}\left(u_i|x_{2,i};\theta_2\right)$$

where x_1, x_2 may or not have common elements.

The above model has a number of desirable attributes. It is theoretically grounded on a distance function and it admits separate explanatory variables for both the mean frontier and the inefficiency term. It also admits a wide variety of distributional assumptions (subject to model-specific identification issues which must be addressed on a case-by-case basis). Importantly, the convolution model measures inefficiency in physical units, unlike other existing models for discrete output which, as will be seen below, measure inefficiency in percentage points (as a marginal effect).

Fé (2007, 2013) explores the use of the Delaporte family of models at length. The Delaporte distribution (Ruohonen 1988) is the convolution of a negative binomial and Poisson distributions. In addition to allowing variations in y due to inefficiency and inputs, the Delaporte family of models also allows variation in y due to unobserved heterogeneity through the over-dispersion parameter in the negative binomial part of the model (see Cameron and Trivedi 1998). By assigning a negative binomial distribution to either y^* or u, researchers can control which term is responsible for heterogeneity in the sample. Inference in this setting is straightforward. The usual maximum likelihood trinity (the score, Wald and likelihood ratio tests) can be used to evaluate the presence of inefficiency or heterogeneity. Different specifications can be tested against each other using Vuong's non-nested hypothesis test (Vuong 1989), whereas the overall adequacy of any model can be tested using the consistent nonparametric extension of Pearson's test for discrete distributions discussed in Fé (2013).

Average inefficiency in the convolution model can be estimated from θ. In the particular case of the Delapote model, the conditional mean of the distribution of u provides a direct estimate of inefficiency in physical units. Following Jondrow et al. (1982), researchers can also obtain point estimates (in physical units) of inefficiency by estimating $E(u_i|y_i, x_i, \theta)$—which in the case of the Delaporte distribution has a closed-form expression.

The convolution model is only suitable for the analysis of economic bads. Hofler and Scrogin (2008) have suggested the use of the Beta-Binomial model for under-reported counts as a candidate CDSF for economic goods. As in the Delaporte family of models, the building block of the model in Hofler and Scrogin (2008) is a stochastic frontier with a negative binomial distribution. Unlike in the Delaporte models, however, the number of units effectively produced, y, given the frontier level y^* is seen as the result of y^* Bernoulli production trials. With probability p, a firm or individual successfully produces the next unit of output, where p is a random variable with a beta distribution. The ensuing unconditional distribution for y is known as a beta-binomial distribution (Fader and Hardie 2000).

The main drawback of this earliest literature is that the convolution and Beta-Binomial models are only applicable to either economic bads or economic goods. However, there is the expectation implicit in the stochastic frontier literature that a SFM will adapt to economic goods and bads alike with a simple transformation in the sign of the inefficiency term. From that perspective, the models in Hofler and Scrogin (2008) and Fé (2007, 2013) are incomplete SFM.

The first complete discrete outcome SFM was proposed by Fé and Hofler (2013). These authors suggested to embed a multiplicative distance function similar to those used in continuous output models into the conditional mean of a Poisson process. Fé and Hofler (2013) define a conditional latent production function $\ln y^* = h(x; \beta) \pm \varepsilon$, where ε is a non-negative constant introducing inefficiency in the process. Then, the number of units produced, y, is assumed to have a conditional density given by $f(y|x, \varepsilon) = \text{Poisson}(\lambda)$ where $\lambda = E(y|x, \varepsilon) = \exp(h(x; \beta) \pm \varepsilon)$. As in Aigner et al. (1977), Fé and Hofler assume that the inefficiency term ε is a random variable with a half-normal distribution. It then follows that the unconditional distribution of y is a Mixed Poisson model with log-half-normal mixing parameter (PHN hereafter),

$$f(y|x) = \int f(y|x; \varepsilon) f(\varepsilon) \mathrm{d}\varepsilon = \int \text{Poisson}(\lambda) f(\varepsilon) \mathrm{d}\varepsilon = E_\varepsilon[\text{Poisson}(\lambda)].$$

This distribution does not have a closed-form expression; however, it can be approximated by simulation. Fé and Hofler construct the simulated likelihood of the model using low discrepancy Halton sequences and provide expressions for the gradient and first moments of the distribution. Testing the significance of inefficiency in this model amounts to applying a score/likelihood ratio/Wald test of zero variance in the half-normal distribution associated with inefficiency. As in Fé (2013), Fé and Hofler propose a consistent conditional moment test for the overall adequacy of the PHN. Unlike the convolution and BB models, the PHN does not accommodate unobserved heterogeneity, however, as shown in Drivas et al. (2014), this is a straightforward extension by defining a three error component model where $\ln y^* = h(x;\beta) \pm \varepsilon + v$ and v is a zero-mean normally distributed error term accounting for any unobserved heterogeneity. Count data are pervasive in economics and social sciences; however, researchers are very often interested in binary outcomes indicating the presence of a characteristic or the attainment of a specific policy goal (such as employment status or the reduction of deficit). To account for inefficiencies in this kind of setting, Ghadge (2017) has proposed an extension of the logit model which has a binary choice stochastic frontier interpretation. The motivation behind his study is understanding the production of intellectual property among firms. Using the popular patent data set in Wang et al. (1998), Ghadge separates firms into those with at least one patent and those with no patents. As in the preceding PHN model, Ghadge suggests a latent production frontier $y^* = h(x'\beta) \pm \varepsilon + v$ where ε is a half-normal inefficiency term and v is the usual zero-mean, unit-variance error component. It is easy to show (using the latent utility-function specification underlying the classic logit model) that y has a Bernoulli distribution conditional on x and ε. The unconditional distribution for the logit stochastic frontier can then be obtained by integration,

$$\begin{aligned} f(y|x) &= \int \Lambda(x'\beta \pm \varepsilon)^{y_i}(1 - \Lambda(x'\beta \pm \varepsilon))^{1-y_i} f(\varepsilon)\mathrm{d}\varepsilon \\ &= E_\varepsilon[\text{Bernoulli}(\Lambda)] \end{aligned}$$

where $\Lambda = \Lambda(x'\beta \pm \varepsilon)$ is the cumulative distribution function of a standard logistic random variable. As in Fé and Hofler (2013), the above density function can be approximated using simulation and inferences about the significance of inefficiency can be drawn by testing the magnitude of the variance of the distribution of ε. However, Monte Carlo simulations in Ghadge (2017) suggest that the parameters of the model are estimated imprecisely,

even for moderately large samples—a point that we revisit in the following section.

One final contribution to the literature on discrete frontier models is the article by Griffiths et al. (2014). These authors explored the Bayesian estimation of an ordered logit SFM. As in the preceding contributions, the building block of their model is a latent distance function $y^* = h(x'\beta) \pm \varepsilon + v$ with $v \sim N(0,1)$ and $\varepsilon \sim HN(\sigma_\varepsilon)$. Unlike in the count data and binary choice models, in their model y can take on one among $j = 1, \ldots, J$ ordered categories and so,

$$f(y = j|x, \varepsilon) = \Phi(\mu_j - x'\beta \mp \varepsilon) - \Phi(\mu_{j-1} - x'\beta \mp \varepsilon).$$

The authors then derive the Bayesian posterior distribution of the model and suggest a Gibb's sampler to estimate the parameters of the model. They use their model to estimate inefficiency in self-reported health, as measured by people's responses to a five-point liker scale.

Although the distribution of outcomes can vary across applications, all the preceding models of stochastic frontier (as well as the continuous data models originated from Aigner et al. [1977]) share the characteristic of being mixed probability distributions with a non-negative error component. In certain situations, the ensuing mixed distribution has a closed-form expression, as is the case in Aigner et al. (1977). More often, however, simulation methods need to be used to facilitate estimation of the parameters of the model. This common structural form allows us to propose an encompassing framework for stochastic frontiers from which all the models just described can be derived and where simulation methods can be routinely used to estimate the parameter of the models. We develop this framework in the following section.

3 A General Framework

Consider an scenario where a researcher has a sample of n observations, $(y_i, x_i)_{i=1}^n$ containing information about a set of inputs $x_i \in \mathcal{X} \subseteq \mathbb{R}^K$ and an outcome of interest $y_i \in \mathcal{Y} \subseteq \mathbb{R}$. The outcome could be an economic good, such as the number of patents awarded to a firm, or a bad, such as the number of infant deaths in a county. The set $\mathcal{Y}$ is left unspecified, but in the kind of situation that will interest us $\mathcal{Y}$ will be a finite or countable subset of $\mathbb{R}$. The maintained assumption is that, underlying y, there is a latent

production function[3] $y_i^* = x'\beta \pm \varepsilon$ which determines the optimal amount of output. Here $\beta \in \mathbb{R}^k$ is a vector of constants which measures the relative contribution of each input. The term $x'\beta$ represents the optimum level of production given the levels of x. The term ε represents inefficiency in production. The relationship between y^* and y will be specific for each different application but in the cases that are discussed below, y^* will characterise the conditional mean in the distribution of y. For the time being, it is assumed that the relationship between y^* and y is fully specified.

We can progress towards an estimable model by further defining a probability density function for y conditional on the levels of ε. For example, the original SFM in Aigner et al. (1977) assumes that $f(y|x, \varepsilon) \sim N(x'\beta \pm \varepsilon, \sigma^2)$ for some $\sigma^2 > 0$. In this case, the latent output y^* corresponds to $E(y|x, \varepsilon)$ and the coefficients β_k, $k = 1, \ldots, K$ are the partial derivatives of y with respect to[4] x_k. One can write the full model as $y = x'\beta + v \pm \varepsilon$ where the distribution of $v \pm \varepsilon$ is the convolution of a normal and a half-normal distributions.[5]

Assume for now that a suitable model has been found for the conditional distribution of y, namely $f(y|x, \varepsilon; \theta) = f\left(x'\beta \pm \varepsilon\right)$ where θ is a vector of parameters including β. We next allow the inefficiency term ε to vary across observations. Because ε is not nonparametrically identified from the data alone, the SFM literature has put forward numerous probabilistic models to characterise the behaviour of ε (as a half-normal, truncated normal, gamma or exponential density). We borrow from Aigner et al. (1977) and assume that ε follows a half-normal distribution, so that $\varepsilon = |u|$ with $u \sim N(0, \sigma_u^2)$ and therefore $f(\varepsilon) = f(\varepsilon; \sigma_u)$. As in the early stochastic frontier literature, we further assume that the density function of inefficiency satisfies $f(\varepsilon|x) = f(\varepsilon)$.

Having defined a distribution for ε, obtaining an unconditional distribution for y is conceptually straightforward,

[3]This function is normally derived from a distance function, with y^* typically being the logarithm of a certain latent output. This is only implicit in the ensuing discussion. Similarly, one would have to distinguish between production functions and cost functions. Once again, we ignore this distinction here for simplicity.

[4]In the ALS model, y and x_k are natural logarithms of output and inputs, in which case the β_k can be interpreted as elasticities. This detail is ignored here for clarity.

[5]This notation makes the stochastic nature of the frontier explicit, since it is readily seen that $h(x') + v \sim N(x'\beta; \sigma_v)$. In general, the stochastic nature of the frontier will be implicitly given by the conditional distribution of y given ε.

$$f(y|x;\theta,\sigma_u) = \int f(y|x,\varepsilon;\theta)f(\varepsilon;\sigma_u)\mathrm{d}\varepsilon = \int f(y|x,\pm\sigma_u|s|;\theta)\phi(s)$$
$$= E_u\left[f(y|x,\pm\sigma_u|s|;\theta)\right]$$

where the second equality follows from the change of variable $s = u/\sigma_u$ implying that s is a random variable with a standard normal distribution $\phi(s)$. This model is fully specified up to the unknown parameters $\theta \in \mathbb{R}^p$ and σ_u, which are the object of estimation. The log-likelihood function of the model is given by

$$\ln L(\theta,\sigma_u) = \sum_{i=1}^{n} \ln\left[\int f(y_i|x_i,\pm\sigma_u|s|;\theta)\phi(s_i)\mathrm{d}s_i\right]$$
$$= \sum_{i=1}^{n} \ln E_u\left[f(y_i|x_i,\pm\sigma_u|s|;\theta)\right].$$

and the estimators of θ and σ_u can be obtained by maximising $\ln L$.

The density $f(y|x;\theta,\sigma_u)$ has a closed-form expression in only a few notable cases (such as the normal-half normal model in Aigner et al. (1977). In general, obtaining a closed-form expression is complicated or infeasible. This is, however, a minor limitation, given advances in simulation-based estimation methods proposed over the last thirty years, a case that in the Stochastic Frontier literature, was first made by Greene (2003). Because the distribution of ε has been pre-specified, the integral in $f(y|x;\theta,\sigma_u)$ can be approximated by simulation, by sampling randomly from $\phi(s)$. Then under a suitable law of large numbers, for fixed θ, σ_u

$$\hat{f}(y_i|x_i) = \frac{1}{R}\sum_{r=1}^{n} f\left(y_i|x_i,\pm\sigma_u|\xi_{i,r}|;\theta\right) \to f(y|x;\theta,\sigma_u)$$

where $\xi_{i,r}$ is a random, independent draw from $\phi(s)$. More precisely, Gourieroux et al. (1984a, b) note that the log-likelihood function can be approximated by any of the following criteria:

$$L^I_{R,n}(\theta) = \sum_{i=1}^{n} \ln\left[\frac{1}{R}\sum_{r=1}^{n} f\left(y_i|x_i,\pm\sigma_u|\xi_{i,r}|;\theta\right)\right] \tag{1}$$

$$L^D_{R,n}(\theta) = \sum_{i=1}^{n} \ln\left[\frac{1}{R}\sum_{r=1}^{n} f\left(y_i|x_i,\pm\sigma_u|\xi_{i,r}|;\theta\right)\right] \tag{2}$$

where the subscripts *I*(dentical) and *D*(ifferent) refer to whether the simulated values, ξ, are or not the same across i. The distinction is not trivial for the asymptotic properties of the method. Letting $\eta = \left(\theta', \sigma_u\right)'$, Gourieroux et al. (1984b) show that

1. If $nR^{-1} \to 0$, then $\sqrt{n}\left(\hat{\eta}^I - \eta_0\right) \to N\left(0, \mathcal{I}^{-1}(\eta_0)\right)$,
2. If $\sqrt{n}R^{-1} \to 0$, then $\sqrt{n}\left(\hat{\eta}^D - \eta_0\right) \to N\left(0, \mathcal{I}^{-1}(\eta_0)\right)$

where $\mathcal{I}(.)$ is the information matrix and η_0 is the true parameter value. Since the number of Monte Carlo draws, R, is decided a priori by the researcher, arbitrary accuracy can be attained with each method by choosing R large enough. Although random sampling with replacement is straightforward with modern computers, that standard Monte Carlo draws tend to form clusters and leave unexplored areas in the unit cube, thus reducing the accuracy of the maximum simulated likelihood method. Low discrepancy sequences (e.g. Halton 1964) give a better coverage and, as with antithetic draws, they generate negatively correlated nodes which result in a reduction of the error due to simulation.[6] For the type of model under consideration, we follow Greene (2003) and Fe and Hofler (2013) and recommend the combination of maximum simulated likelihood with Halton sequences (Halton 1964).[7]

The analytical derivatives of the simulated likelihood function are given by

$$\frac{\partial L^D_{R,n}(\eta)}{\partial \eta} = \sum_{i=1}^{n} \hat{g}_i(\eta) = \sum_{i=1}^{n} \frac{1}{\hat{f}(y_i|x_i)} \frac{1}{R} \sum_{r=1}^{R} \frac{\partial f\left(y_i|x_i, \pm\sigma_u|\xi_{i,r}|; \theta\right)}{\partial \eta}.$$

and similarly for $\partial L^I_{R,n}(\eta)/\partial\eta$. The value $\hat{\theta}_{MSL}$ making the above system of equations equal to zero is the maximum simulated likelihood estimator of θ. When the model for $f(y|x; \theta, \sigma_u)$ is correctly specified, the $\hat{\eta}_{MSL}$ is a

[6]This occurs because the variance of the sum of any two draws is less than the variance of the sum of two independent draws (Gentle 2003). A clear drawback of Quasi-Monte Carlo methods is the deterministic nature of the sequence which results in negatively correlated random draws (even though the strength of the correlation might be small). This contradicts the assumptions of independent random draws on which the above asymptotic results hold. It is possible, however, to generate randomised Halton draws without compromising the properties of the original sequence. In particular, Bhat (2003) advocates shifting the original terms of the Halton sequence, say s_h, by a quantity μ that has been drawn randomly from a uniform distribution on [0, 1]. In the resulting sequence, those terms exceeding 1 are transformed so that $s^*_h = s_h + \mu - 1$.

[7]Halton sequences are now well established in applied econometrics and therefore we refer the reader to excellent surveys by, among others, Niederreiter (1992), Gentle (2003) and Train (2003).

consistent, efficient estimator of θ and its variance can be calculated using the outer product of gradients,

$$\hat{V}(\eta_{MSL}) = \left(\sum_{i=1}^{n} \hat{g}_i\left(\hat{\eta}_{MSL}\right) \hat{g}_i\left(\hat{\eta}_{MSL}\right)' \right)^{-1}.$$

Furthermore, under no misspecification, $\hat{\eta}_{MSL}$ has an asymptotic normal distribution, which readily enables the construction of conditional moment tests. In particular, evaluating the existence or magnitude of inefficiency involves a Wald, Score or Likelihood Ratio tests of $H_0 : \sigma_u = 0$. These tests will follow a chi-square with degrees of freedom equal to the number of restrictions.

The parameters in θ will often be of interest in their own right. However, the ultimate goal of the analysis is to provide average and point estimates of the level of inefficiency. This can be done by following Jondrow et al. (1982) and applying Bayes' theorem to obtain cross-sectional inefficiency scores for ε through $E(\varepsilon|y,x) = \int \varepsilon f(\varepsilon|x,y)\mathrm{d}\varepsilon$, where

$$f(\varepsilon|x,y) = \frac{f(y|x,\varepsilon;\theta)f(\varepsilon)}{f(y|x)} = \frac{f(y|x,\pm\sigma_u|s|;\theta)\phi(s)}{f(y|x)}.$$

When $E(\varepsilon|y,x) = \int \varepsilon f(\varepsilon|x,y)\mathrm{d}\varepsilon$ fails to have a closed-form expression, it can be approximated via simulation[8]

$$\hat{\varepsilon}_i = E(\varepsilon_i|x_i,y_i) \approx \frac{1}{R}\sum_{r=1}^{R} \frac{\pm\sigma|\xi_{i,r}|f\left(y_i|x_i,\pm\sigma_u|\xi_{i,r}|;\theta\right)}{\hat{f}(y_i|x_i)}. \tag{3}$$

The general framework in the preceding discussion can be applied to a number of models. Next, we illustrate this by introducing count, binary and ordered outcome SFM.[9]

[8]Wang and Schmidt (2009) observe that the distributions of ε and $\hat{\varepsilon}$ are not the same. In particular, $\hat{\varepsilon}$ has smaller variance and the lower and upper tails of the distribution of ε will be misreported, so that $\hat{\varepsilon}$ penalises outstanding firms and rewards the least efficient individuals—although the average efficiency in the sample is correctly approximated. Quoting Wang and Schmidt (2009), this does not mean that there is anything wrong with the estimator since it is unbiased in the unconditional sense $E(\hat{\varepsilon} - \varepsilon) = 0$.

[9]The codes to implement these models (written in Ox 7.10, Doornik [2011]) are available from the author upon requests.

3.1 Count Data Stochastic Frontier

The CDSF was first discussed by Fé and Hofler (2013), and therefore we provide here a brief outline of the model. The starting point is the conditional frontier $\ln y^* = x'\beta \pm \varepsilon$ which determines the mean of y insofar $E(y|x;\varepsilon) = \lambda = \exp(x'\beta \pm \varepsilon) > 0$. Since y is discrete-valued over the set of non-negative integers, Fé and Hofler (2013) follow the convention and assume that the stochastic frontier is distributed $f(y|x,\varepsilon) = \text{Poisson}(\lambda)$. and therefore, y has a Poisson Log-Half-Normal (PHN) distribution,

$$f(y|x;\beta,\sigma_u) = \int \text{Poisson}(\lambda) f(\varepsilon) \mathrm{d}\varepsilon = E_u\left[\text{Poisson}\left(\exp\left(x'\beta \pm \sigma_u |s|\right)\right)\right]$$

where $s \sim N(0,1)$. This distribution can be approximated by simulation as described above, by drawing from the standard normal distribution.

3.2 Binary Choice Stochastic Frontier Model

Researchers are often interested in explaining binary outcomes indicating the achievement of goals or the presence of attributes. For example, economists have investigated the effectiveness of firms' expenditure in research and development and whether or not this results in the award of patents. Labour economists often report the results of randomised interventions where the outcome of interest is a person's employment status or the achievement of a certification. Investments in these activities are costly, and there is a case to try to understand the efficiency of these investments. This can be done by extending popular binary choice models, such as the Probit and Logit models, to a Stochastic Frontier setting. This has been proposed by Ghadge (2017), who used a mixed Bernoulli distribution, with a non-negative mixing variable, to obtain an estimable Logit SFM. In this section, we propose an alternative approach which is simpler to implement in practice and is transparent about the identification conditions that the model must satisfy.

The point of departure is the latent stochastic frontier $y^* = x'\beta \pm \varepsilon + v$. The outcome of interest, y takes on values 0 or 1, depending on whether a specific outcome has occurred. In particular,

$$y = \begin{cases} 1 & \text{if } y^* > 0 \equiv -x'\beta \mp \varepsilon < v \\ 0 & \text{if } y^* \leq 0 \equiv -x\beta \mp \varepsilon \geq v \end{cases} \tag{4}$$

As in the standard binary choice literature, we assume that (i) $E(u|x) = E(v|x) = 0$, (ii) $\text{cov}(u, v) = 0$, (iii) $\text{var}(v) = 1$ and (iv) $\text{var}(u) = \sigma_u^2$, where $\varepsilon = |u|$. Assumption (i) rules out endogenous regressors, assumption; (ii) imposes that the error and inefficiency terms be uncorrelated and assumption; (iii) is a standard normalisation required for identification.

Unlike in the standard binary choice literature, the inclusion of inefficiency introduces a problem of identification if x contains an intercept. Conditional of the value of ε, Eq. (4) can be written as

$$y = \begin{cases} 1 & \text{if } x'\beta + v > \mp\varepsilon \\ 0 & \text{if } x\beta + v < \mp\varepsilon \end{cases}$$

where inefficiency is shifting the decision threshold. Let $x = (1, z)'$ and $\beta = (\alpha, \phi)$ where α denotes the coefficient of an intercept. Then,

$$P\left(y^* > \mp\varepsilon\right) = P(\alpha + z'\phi + v > \mp\varepsilon) = P((\alpha \pm \varepsilon) + z'\phi + v > 0)$$

where $(\alpha \pm \varepsilon)$ remains unknown. Data do not contain information to separate α and ε. In practice, then, we must introduce the normalisation $\alpha = 0$ and treat the inefficiency term as a (unconditional) random intercept. With this normalisation, and letting $v \sim N(0, 1)$, we have

$$P(y = 1|x, \varepsilon) = \Phi(x'b \pm \varepsilon) \text{ and } P(y = 0|x, \varepsilon) = \Phi(-x'b \pm \varepsilon)$$
$$\Rightarrow P(y|x, \varepsilon) = \Phi(q(x'\beta \pm \varepsilon))$$

where $q = 2y - 1$. Note that, so far, the discussion falls within the general framework presented at the beginning of the section, insofar we are only specifying the mapping between y and y^* and the stochastic frontier, given by the conditional distribution of y given x and ε.

The above is a random coefficients probit model, where the unconditional distribution of y is

$$P(y = 1|x) = \int \Phi\left(q(x'\beta \pm \varepsilon)\right) f(\varepsilon)\mathrm{d}\varepsilon = \int \Phi(q(x'\beta \pm \sigma_u|u|))\phi(u)\mathrm{d}u$$

and $\phi(u)$ is a standard normal density. This integral does not have a closed-form expression, but it can be approximated by simulation,

$$L_{R,n}^{D}(\theta) = \sum_{i=1}^{n} \ln \left[\frac{1}{R} \sum_{r=1}^{R} \Phi\left(q\left(x'\beta \pm \sigma |\xi_{i,r}|\right)\right) \right] \tag{5}$$

where $\xi_{i,r}$ $i = 1, \ldots, n$, $r = 1, \ldots, R$ are draws from the standard normal distribution. The principal hypothesis of interest in this model is $H_0 : \sigma = 0$, in which case the model collapses to a standard probit model without an intercept. To obtain point estimates of inefficiency, it suffices to apply definition (3) combined with (5).

3.3 Ordered Logit Stochastic Frontiers

The preceding model can be extended to situation where y is an ordered categorical variable. This extension has been considered by Griffiths et al. (2014) from a Bayesian perspective. The motivation behind the original paper by Griffiths et al. (2014) was the estimation of a production function for health (measured in a 5-point Likert scale). Here we present the *frequentist* version of the model. As before, the starting point is the latent production frontier $y^* = x'\beta \pm \varepsilon + v$ where v has a standard normal distribution. For similar reasons to those seen in the probit model, now x cannot include an intercept term. The ordinal variable y is such that

$$y = \begin{cases} 0 & \text{if } y^* \leq 0 \\ 1 & \text{if } 0 < y^* \leq \mu_1 \\ 2 & \text{if } \mu_1 < y^* \leq \mu_2 \\ \vdots & \\ J & \text{if } \mu_{J-1} \leq y^* \end{cases}$$

The μ_j, $j = 1, \ldots, J-1$ are threshold parameters which need to be estimated alongside β. As in the probit model, the inefficiency term ε shifts the thresholds to left or right, thus determining the degree of skewness in the distribution of y. Under the distributional assumptions above, it follows that

$$\begin{gathered} f(y = 0|x, \varepsilon) = \Phi(-x'\beta \mp \varepsilon) \\ f(y = 1|x, \varepsilon) = \Phi(\mu_1 - x'\beta \mp \varepsilon) - \Phi(-x'\beta \mp \varepsilon) \\ f(y = 2|x, \varepsilon) = \Phi(\mu_2 - x'\beta \mp \varepsilon) - \Phi(\mu_1 - x'\beta \mp \varepsilon) \\ \vdots \\ f(y = J|x, \varepsilon) = 1 - \Phi(\mu_{J-1} - x'\beta \mp \varepsilon) \end{gathered}$$

where $0 < \mu_1 < \mu_2 < \ldots < \mu_{J-1}$. Integrating ε out results in the unconditional probabilities

$$f(y=j|x) = \int \Phi\big(\mu_j - x'\beta \mp \sigma|u|\big)\phi(u)\mathrm{d}u - \int \Phi\big(\mu_{j-1} - x'\beta \mp \sigma|u|\big)\phi(u)\mathrm{d}u$$

Letting $z_j = \mathbb{I}(y=j)$ we have

$$\begin{aligned} f(y|x) &= \prod_{j=1}^{J} f(y=j|x)^{z_j} \\ &= \prod_{j=1}^{J} \left[\int \Phi\big(\mu_j - x'\beta \mp \sigma|u|\big)\phi(u)du - \int \Phi\big(\mu_{j-1} - x'\beta \mp \sigma|u|\big)\phi(u)du \right]^{z_j} \end{aligned}$$

As before, this distribution can be approximated by simulation,

$$\hat{f}(y|x) = \prod_{j=1}^{J} \left[\frac{1}{R}\sum_{r=1}^{R} \Phi\big(\mu_j - x'\beta \mp \xi_{i,r}|u|\big) - \frac{1}{R}\sum_{r=1}^{R} \Phi\big(\mu_{j-1} - x'\beta \mp \xi_{i,r}|u|\big) \right]^{z_j}$$

in which case estimation of the parameters of the model follows by maximising the simulated likelihood,

$$L_{R,n}^{D}(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{J} z_j \ln \left[\frac{1}{R}\sum_{r=1}^{R} \Phi\big(\mu_j - x'\beta \mp \xi_{i,r}|u|\big) - \frac{1}{R}\sum_{r=1}^{R} \Phi\big(\mu_{j-1} - x'\beta \mp \xi_{i,r}|u|\big) \right]$$

4 Extensions

4.1 Unobserved Heterogeneity

In a series of influential articles, Greene (2004a, 2005) notes the importance of separating inefficiency from unobserved heterogeneity. The latter could be due to variation in socio-economic, personality or cultural characteristics. Ignoring its presence would result in biased estimates of inefficiency and, more generally, the parameters of the model.

In a cross-sectional setting, the preceding models can be adapted to take into account unmeasured heterogeneity. Drivas et al. (2014) proposed such an extension for the CDSF model in Fé and Hofler (2013). Their extension consists in adding a second error component in the conditional mean of the count variable y, so that $E(y|x,\varepsilon) = \exp(x'\beta \pm \varepsilon + v)$, where $v \sim N(0,\sigma_v^2)$.

A similar extension for the probit stochastic frontier models can be obtained by defining

$$P(y|x,\varepsilon,w) = \Phi\big(q(x'\beta \pm \varepsilon + w)\big)$$

where w is a zero-mean, normal random variable such that $f(w|\varepsilon) = f(w)$. Unlike in the CDSF model, we need to set $\sigma_w = 1$ for identification. The unconditional distribution of y is given by

$$P(y|x) = \int_0^{\infty}\int_{-\infty}^{\infty} \Phi(q(x'\beta \pm \varepsilon + w))f(\varepsilon)\phi(w)\mathrm{d}\varepsilon\mathrm{d}w$$

which can be approximated by simulation,

$$\hat{P}(y|x) = \frac{1}{R}\frac{1}{R'}\sum_{r=1}^{R}\sum_{r'=1}^{R'} \Phi\big(q\big(x'\beta \pm \sigma_u|\xi_r| + \zeta_{r'}\big)\big)$$

where ξ, ζ are draws from a standard normal distribution. This approximation can then be used as the kernel in the simulated likelihood (5).

To study the effect of heterogeneity in the estimation of the probit stochastic frontier, we run a limited Monte Carlo simulation. We produced 500 data sets from a probit model with $y^* = x_1 + x_2 + v - \sigma_u|u| + w$, where u, v and w are $N(0,1)$,$\sigma_u = 0.5$ and $x_1 \sim U[-1,1]$, $x_2 \sim U[-1,1]$ have a correlation coefficient of 0.5. Here w corresponds to the unobserved heterogeneity in the sample. We estimated correctly specified probit SFM

Table 1 Monte Carlo simulation. Data were generated from a probit model with conditional mean $y^* = x_1 + x_2 + v - \sigma_u|u| + w$, where $\sigma = 0.5$, $v \sim w \sim N(0, 1)$, $u \sim N(0, \sigma^2)$. We estimated the Probit SFM without and with a heterogeneity error component (Models 1 and 2, respectively). The sample size was 500 and the number of terms in the Halton sequences was obtained by setting $R = R' = 50$. We report the mean estimate, bias and mean squared error obtained from 1000 replications

	Model 1			Model 2		
	$\sigma = 0.5$	$\beta_1 = 1$	$\beta_2 = 1$	$\sigma = 0.5$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	0.389	0.740	0.704	0.552	1.019	0.965
Bias	−0.111	−0.260	−0.296	0.052	0.019	−0.035
MSE	0.052	0.266	0.142	0.079	0.375	0.104

without unobserved heterogeneity (Model 1) and with unobserved heterogeneity (Model 2). The number of Halton terms used to simulate the likelihood function were determined by setting $R = R' = 50$. The sample size was set at 500 observations.

Table 1 collects the results. The main feature of the table is the large, downward bias observed in the estimates reported when using Model 1. This is particularly troubling for the inefficiency parameter, σ_u. When using Model 2, as expected, we obtain accurate estimates of the true parameters. Given the small biases associated with this model, we can explain the larger mean square errors by an increase in the variance resulting from both the additional heterogeneity parameter and the additional simulation error due to the second error component.

4.2 Local Likelihood Estimation

The discrete outcome stochastic frontiers discussed in the preceding section rely on parametric assumptions regarding the distribution of inefficiency and the conditional distribution of the outcome. The magnitude of the set of assumptions is significant and, therefore, researchers might question their suitability, primarily because maximum likelihood estimates based on a misspecified model are consistent for a pseudo-true parameter value which might differ from the parameter of interest for the researcher.

Smoothing techniques, such as the Local Linear Regression (REF) can be used to mitigate the effects of model misspecification, but in the current context it is customary to accommodate the under-/over-production induced by inefficiency in the data. This can be done vial Local Maximum Likelihood (Tibshirani and Hastie 1987), a method that allows the parameters of the model (but not necessarily the type of distribution itself) to vary along the domain of x, so that $\theta = \theta(x)$. For example, in the normal

regression case this would imply that the conditional distribution of y at the value x follows a $N\left(\mu(x), \sigma^2(x)\right)$ distribution.

For simplicity, consider here the case where $\theta(x) = (\theta_1(x), \theta_2(x), \ldots, \theta_P(x))'$ but $X \in \mathbb{R}$.[10] Let $\tilde{\theta}(x) = (\tilde{\theta}_1(x), \ldots, \tilde{\theta}_P(x))$ with $\tilde{\theta}_p(x) = \theta_p^{(0)}(x) + \theta_p^{(1)}(x)(x_i - x) + \cdots + \theta_p^{(M)}(x)(x_i - x)$ and let $\Theta = \left(\theta_1^{(0)}(x), \ldots, \theta_P^{(M)}(x)\right)$ be the $(P \times (M+1)) \times 1$ vector containing all the $\theta_p^{(m)}$. The LML estimator for each parameter $\theta_p(x)$ is the value $\hat{\theta}_p^{(0)}(x)$ for $p = 1, \ldots, P$ that solves,

$$\max_{\Theta} \sum_{i=1}^{n} \log L(\theta(x)) K_h(X_i - x) \tag{6}$$

where, $K_h(X_i - x) = h^{-1}k(x_i - x/h)$, $k(.)$ is a univariate kernel function and $h_j = h_j(n)$ is a smoothing parameter such that $h \to 0$ as $n \to \infty$. Tibshirani and Hastie (1987) and Kumbhakar et al. (2007) show that $\hat{\theta}_0(x)$ is a consistent, asymptotic normal estimator of $\theta_0(x)$ as $\sqrt{nh} \to \infty$.

For the general SFM model in Eq. (6), we can replace $L(.)$ with its simulated version, which would result in the parameters of interest are $\theta(x) = (\mu(x), \sigma(x))'$ in the conditional distribution

$$\mathbb{P}(y|x; \theta(x)) = \int_{\varepsilon} \text{Poisson}(\exp(\mu(x) \pm \sigma(x)|u|) f(\varepsilon) \mathrm{d}\varepsilon \tag{7}$$

The log-link in the conditional mean is present to guarantee the non-negativity of the mean parameter but, as we will show in the simulations, it turns out to be a rather innocuous assumption. The infeasible local linear $(m = 1)$ conditional maximum likelihood function is

$$\mathbb{P}(y_i|x_i; \theta(x)) = \sum_{i=1}^{n} \left[\log \int_{u} \text{Poisson}\{\lambda_i(x; u_i)\} f(u_i) du_i\right] K_H(X_i - x) \tag{8}$$

[10]The case of multivariate X is a straightforward extension and is discussed in Tibshirani and Hastie (1987).

where

$$\lambda_i(x;u_i) = \exp\Big(\mu^{(0)}(x) + \mu^{(1)}(x)(X_i - x)' \pm \Big(\sigma^{(0)}(x) + \sigma^{(1)}(x)(X_i - x)'\Big)|u_i|\Big). \tag{9}$$

Local likelihood functions of higher order than one can be obtained similarly. As in the parametric case, the integral in (8) can be approximated by simulation, yielding the simulated local likelihood function, which can then be implemented with standard econometric software.

For the nonparametric model, the appropriate local estimator of inefficiency is

$$\hat{v}_i(X_i) = \approx \frac{\sum_{h=1}^{H} e^{\pm|s_h|\sigma^{(0)}}\text{Poisson}\big(\exp\big(\mu^{(0)}(X_i) + \sigma^{(0)}(X_i)|s_h|\big)\big)}{\sum_{h=1}^{H}\text{Poisson}\big(\exp\big(\mu^{(0)}(X_i) + \sigma^{(0)}(X_i)|s_h|\big)\big)} \tag{10}$$

Feasible estimators would follow by using the (local) maximum simulated likelihood estimates of β, σ, $\mu^{(0)}(x)$ and $\sigma^{(0)}(x)$. Nonparametric models may fit the data better, but it must be noted that they may provide limited information as to the shape of frontier. Thus, it may be difficult to gauge whether the nonparametric model fits with economic expectations. As such there is still a role for the parametric model, and the tests for its adequacy presented in the previous sections are bound to be useful in this respect.

To assess the performance of this estimator, we run a limited Monte Carlo simulation. Data were drawn 1000 times from a Poisson Half-Normal distribution with conditional mean $\lambda = \exp(1 + x - \sigma_u|s|)$ where $\sigma_u = 1$, $s \sim N(0,1)$ and $x \sim U[-1,1]$. In each replication, we estimated the mean and σ_u of the model at $x = \pm 0.77$ and $x = \pm 0.33$ using the local likelihood estimator with $m = 0$ and bandwidth $h = 2 * \sigma_x N^{1/5}$. We considered three sample sizes, $N = 100$, 500, 1000. The mean estimate, bias and the standard deviation of the simulation are reported in Tables 2 and 3. The LML estimator works well for both the mean and the inefficiency parameters. The average bias is small, even when $N = 100$ and it tends to decrease with the sample size (as does the standard deviation). The parameters tend to be estimated slightly less precisely at the right tail of the distribution of x.

Table 2 Local likelihood estimation of the mean in the CDSF, based on 1000 Monte Carlo Replications. Bandwidth $=2\sigma_x N^{-1/5}$

x	Mean	True	Bias	S.D.
$N = 100$				
−0.77	1.25	1.25	0.00	0.80
−0.33	1.89	1.95	−0.06	0.97
0.33	3.73	3.79	−0.06	1.33
0.77	5.81	5.92	−0.11	1.69
$N = 500$				
−0.77	1.21	1.25	−0.04	0.59
−0.33	1.88	1.95	−0.06	0.74
0.33	3.75	3.79	−0.07	0.94
0.77	5.91	5.92	−0.02	1.12
$N = 1000$				
−0.77	1.19	1.25	−0.05	0.52
−0.33	1.88	1.95	−0.07	0.65
0.33	3.75	3.79	−0.04	0.78
0.77	5.90	5.92	−0.01	0.91

Table 3 Local likelihood estimation of the inefficiency parameter σ_u in the CDSF, based on 1000 Monte Carlo Replications. Bandwidth, $h = 2\sigma_x N^{-1/5}$

x	Mean	True	Bias	S.D.
$N = 100$				
−0.77	1.06	1	0.06	1.52
−0.33	0.92	1	−0.08	0.92
0.33	0.97	1	−0.03	0.63
−0.77	0.98	1	−0.02	0.51
$N = 500$				
−0.77	0.92	1	−0.07	0.92
−0.33	0.94	1	−0.06	0.71
0.33	0.98	1	−0.02	0.45
−0.77	1.02	1	0.02	0.32
$N = 1000$				
−0.77	0.92	1	−0.08	0.81
−0.33	0.93	1	−0.07	0.63
0.33	1.01	1	0.01	0.36
−0.77	1.03	1	0.03	0.27

5 Conclusion

This chapter reviews recent contributions to the area of SFM for the analysis of discrete outcomes. More specifically, we discuss models for binary indicators (probit SFM), ordered categorical data (ordered logit SFM) and discrete outcomes (Poisson SFM).

All these models are mixtures of a standard distribution with an asymmetric distribution. This allows us to frame the discussion within a general framework from which most SFM can be derived. Because many of these models might lack a closed-form likelihood function, we suggest the use of maximum simulated likelihoods to estimate the parameters of each model. The latter method is easy to implement in a modern computer and the unknown likelihood can be approximated with arbitrary accuracy using low discrepancy sequences such as Halton sequences.

The construction and estimation of SFM for discrete data are a relatively new area of research. It is, however, one with great potential for applications, especially in areas such as health, labour and industrial economics, where the outcomes of interest are often discretely distributed. These models are also applicable to the related problem of under-reporting in survey data. In criminology, for example, police data are routinely used to study violent or anti-social behaviour. However, it is well documented that certain types of crimes (domestic violence in particular), are too often not reported to the police. The models in this survey can be applied to those situations, with the inefficiency term estimating the magnitude of under-reporting in the data.

This survey has not covered models for longitudinal data and, throughout the discussion, it has been assumed that inefficiency is exogenous to inputs and outputs. Random effects discrete SFM for longitudinal data are a straightforward modification of the models surveyed in this chapter. Fixed effects versions of these models, however, will need to address the pervasive incidental parameter problem (Neyman and Scott 1948; Lancaster 2000). Relaxing the implicit assumption of exogenous inefficiency implies an even greater challenge, requiring the identification of instrumental variables, as well as a theoretically justified estimation methods. Neither of these challenges seem, however, insurmountable given recent advances seen in the areas of bias correction in non-linear fixed effects models and the instrumental variable estimation of non-linear models. Addressing these challenges is, however, left for future research.

Acknowledgements The author thanks Richard Hofler and William Greene for helpful comments.

References

Aigner, D., C. Lovell, and P. Schmidt. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6: 21–37.

Angrist, J., and J. Pischke. 2008. *Mostly harmless econometrics.* Princeton: Princeton University Press.

Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio Codina. 2015. Estimating the production function for human capital: results from a randomized controlled trial in Colombia. IFS Working Paper W15/06, Institute for Fiscal Studies.

Ben-Porath, Y. 1967. The production of human capital and the life cycle of earnings. *Journal of Political Economy* 75 (4): 352–365.

Bhat, C. 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B (Methodological)* 37: 837–855.

Cameron, C., and P. Trivedi. 1998. *Regression analysis of count data*. Cambridge: Cambridge University Press.

Chamberlin, E.H. 1933. *Theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

Cobb, C.H., and P.H. Douglas. 1928. A theory of production. *American Economic Review* 18: 139–165.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Doornik, J. A. 2011. *Ox: An object-oriented matrix programming language*. London: Timberlake Consultants Ltd.

Drivas, K., C. Economidou, and E. Tsionas. 2014. A Poisson stochastic frontier model with finite mixture structure. Mpra paper, University Library of Munich, Germany.

Fader, P., and B. Hardie. 2000. A note on modelling underreported Poisson counts. *Journal of Applied Statistics* 8: 953–964.

Fé, E. 2007. Exploring a stochastic frontier model when the dependent variable is a count. Economics discussion EDP-0725, University of Manchester.

Fé, E. 2013. Estimating production frontiers and efficiency when output is a discretely distributed economic bad. *Journal of Productivity Analysis* 39 (3): 285–302.

Fé, E., and R. Hofler. 2013. Count data stochastic frontier models, with an application to the patents-R&D relationship. *Journal of Productivity Analysis* 39 (3): 271–284.

Gentle, J. E. 2003. Random number generation and Monte Carlo methods, 2nd ed. New York: Springer.

Ghadge, C.A. 2017. Contributions to the inference on stochastic frontier models. Ph.D. dissertation, Savitribai Phule Pune University.

Gourieroux, C., A. Monfort, and A. Trognon. 1984a. Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica* 52: 701–720.

Gourieroux, C., A. Monfort, and A. Trognon. 1984b. Pseudo maximum likelihood methods: Theory. *Econometrica* 52: 681–700.

Greene, W.H. 2003. Simulated likelihood estimation of the normal-gamma stochastic frontier function. *Journal of Productivity Analysis* 19 (2): 179–190.

Greene, W. 2004a. Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Economics* 13: 959–980.

Greene, W. 2004b. *Econometric analysis*. Upper Saddle: Prentice Hall.

Greene, W. 2005. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics* 126: 269–303.

Greene, W., and D. Hensher. 2010. *Modeling ordered choices: A primer*. Cambridge: Cambridge University Press.

Griffiths, W., X. Zhang, and X. Zhao. 2014. Estimation and efficiency measurement in stochastic production frontiers with ordinal outcomes. *Journal of Productivity Analysis* 42 (1): 67–84.

Grossman, M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 82: 233–255.

Hall, B., Z. Griliches, and J.A. Hausman. 1986. Patents and R&D: Is there a lag? *International Economic Review* 27: 265–283.

Halton, J.H. 1964. Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM* 7 (12): 701–702.

Heckman, J.J. 1976. A life-cycle model of earnings, learning, and consumption. *Journal of Political Economy* 84 (4): S9–S44.

Hicks, J.R. 1935. Annual survey of economic theory: The theory of monopoly. *Econometrica* 3 (1): 1–20.

Hofler, R., and D. Scrogin. 2008. A count data stochastic frontier. Discussion paper, University of Central Florida.

Hottenrott, H., B.H. Hall, and D. Czarnitzki. 2016. Patents as quality signals? The implications for financing constraints on R&D. *Economics of Innovation and New Technology* 25 (3): 197–217.

Jondrow, J., I. Materov, K. Lovell, and P. Schmidt. 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics* 19: 233–238.

Koopmans, T. 1951. An analysis of production as an efficient combination of activities. Monograph 13, Cowles Commission for Research in Economics.

Kumbhakar, S., B. Park, L. Simar, and E. Tsionas. 2007. Nonparametric stochastic frontiers: A local maximum likelihood approach. *Journal of Econometrics* 137: 1–27.

Lancaster, A. 2000. The incidental parameter problem since 1948. *Journal of Econometrics* 95: 391–413.

Meeusen, W., and J. van den Broeck. 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* 18: 435–444.

Mincer, J. 1981. Human capital and economic growth. Technical report, National Bureau of Economic Research.

Neyman, J., and E. Scott. 1948. Consistent estimation from partially consistent observations. *Econometrica* 16: 1–32.

Niederreiter, H. 1992. *Random number generation and Quasi-Monte Carlo methods*. Philadelphia: Society for Industrial and Applied Mathematics.

Ruohonen, M. 1988. On a model for the claim number process. *ASTIN Bulletin* 18 (1): 57–68.

Shephard, R. (1970). *Theory of cost and production function*. Princeton: Princeton University Press.

Tibshirani, R., and T. Hastie. 1987. Local likelihood estimation. *Journal of the American Economic Association* 82: 559–567.

Train, K. 2003. *Dicrete choice methods with simulation*. Cambridge: Cambridge University Press.

Vuong, Q.H. 1989. Likelihood ratio tests for model selection and non-nested hypothesis. *Econometrica* 57: 307–333.

Wang, W., and P. Schmidt. 2009. On the distribution of estimated technical efficiency in stochastic frontier models. *Journal of Econometrics* 148: 36–45.

Wang, P., I.M. Cockburn, and M.L. Puterman. 1998. Analysis of patent data: A mixed-poisson-regression-model approach. *Journal of Business and Economic Statistics* 16: 27–41.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 53: 1–16.

Wicksteed, P. 1894. *An essay on the coordination of the laws of distribution*. London: Macmillan.

Wildman, J., and A. Jones. 2008. Disentangling the relationship between health and income. *Journal of Health Economics* 27: 308–324.

Nonparametric Statistical Analysis of Production

Camilla Mastromarco, Léopold Simar and Paul W. Wilson

1 Introduction

Benchmarking performance is a common endeavor. In sports, individual athletes as well as teams of athletes strive to win by performing better than their opponents and to set new records for performance. In medicine, physicians and researchers strive to find treatments that enhance or extend patients' lives better than existing treatments. In education, schools attempt to enhance students' prospects for success at the next level or in the workplace. In manufacturing, firms attempt to convert inputs (e.g., land, labor, capital, materials, or energy) into outputs (e.g., goods or services) as "efficiently" (i.e., with as little waste) as possible.

C. Mastromarco
Dipartimento di Scienze dell'Economia, Università del Salento, Lecce, Italy
e-mail: camilla.mastromarco@unisalento.it

L. Simar
Institut de Statistique, Biostatistique, et Sciences Actuarielles,
Université Catholique de Louvain, Louvain-la-Neuve, Belgium
e-mail: leopold.simar@uclouvain.be

P. W. Wilson (✉)
Department of Economics and Division of Computer Science,
School of Computing, Clemson University, Clemson, SC, USA
e-mail: pww@clemson.edu

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_10

This chapter provides an up-to-date survey of statistical tools and results that are available to applied researchers using nonparametric, deterministic estimators to evaluate producers' performances.[1] Previous surveys (e.g., Simar and Wilson 2008, 2013) give summaries of results that allow researchers to estimate and make inference about the efficiency of individual producers. In the review that follows, we describe new results that permit (i) inferences about mean efficiency (both conditional and unconditional), (ii) tests of convexity versus non-convexity of production sets, (iii) tests of constant versus variable returns to scale, (iv) tests of the "separability condition" described by Simar and Wilson (2007) required for second-stage regressions of efficiency estimates on some explanatory variables, and (v) dimension reduction to circumvent the slow convergence rates of nonparametric efficiency estimators. We also show how conditional efficiency estimators can be used when panel data are available to model and detect changes over time.

A rich economic theory of efficiency in production has developed from the pioneering work of Debreu (1951) and Koopmans (1951). Farrell (1957) made the first attempt to estimate a measure of efficiency from observed data on production, but the statistical properties of his estimator were developed much later. Researchers working in economics, econometrics, management science, operations research, mathematical statistics, and other fields have contributed to what has by now become a large literature on efficiency of production.[2]

Measurement of performance (i.e., efficiency) requires a benchmark. A commonly used benchmark is the efficient production frontier, defined in the relevant input-output space as the locus of the maximal attainable level of outputs corresponding to given levels of inputs.[3] As discussed below, it is also possible to measure efficiency against other benchmarks provided by

[1]By using "deterministic" to describe these estimators, we follow the language of the literature. The deterministic estimators do not admit a two-sided noise term, in contrast to the literature on efficiency estimation in the context of parametric, stochastic frontier models. Nonetheless, efficiency must be estimated since, as discussed below, it cannot be observed directly. Even if the researcher has available observations on all existing firms in an industry, he should consider that a clever entrepreneur might establish a new firm that out-performs all of the existing firms. Truth cannot be learned from data.

[2]On January 24, 2018, a search on Google Scholar using the keywords "efficiency," "production," "inputs," and "outputs" returned approximately 2,590,000 results.

[3]Alternatively, if prices of inputs are available, one can consider a cost frontier defined by the minimal cost of producing various levels of outputs. Or if prices of outputs are available, one can consider a revenue frontier defined by the maximal revenue obtained from using various levels of inputs. If both prices of inputs and outputs are available, one can consider a profit frontier. All of these possibilities relate to allocative efficiency.

features that lie "close" to the efficient production frontier. In either case, these benchmarks are unobservable and must be estimated. As such, statistical inference is needed before anything can be learned from data. In turn, a statistical model is needed, as without one inference is not possible as the theoretical results of Bahadur and Savage (1956) make clear.

In the survey that unfolds below, Sect. 2 presents economic and statistical assumptions that comprise a statistical model in which inference about technical efficiency is possible. Section 3 discusses nonparametric estimators of efficiency. Both cases mentioned above are considered, i.e., where either the efficiency frontier or a feature near the efficient frontier is used as a benchmark. Section 4 introduces "environmental" variables that are neither inputs nor outputs, but which may nonetheless affect production. This requires a new statistical model, which is introduced in Sect. 4.1 as well as new estimators which are discussed in Sect. 4.2. Section 5 reviews recent developments providing central limit theorems for mean efficiency when sample means of nonparametric efficiency estimators are used to estimate mean efficiency. Section 6 shows how one can test hypotheses about the structure of the efficient production frontier. Section 7 discusses strategies for increasing accuracy in estimation of efficiency by reducing dimensionality. Section 9 discusses recent developments in dynamic settings, and Sect. 10 provides conclusions.

2 Theory of Production

2.1 Economic Model

Standard economic theory of the firm (e.g., Koopmans 1951; Debreu 1951; or Varian 1978) describes production in terms of a production set

$$\Psi = \left\{(x, y) \in \mathbb{R}_+^{p+q} | x \text{ can produce } y\right\} \tag{1}$$

where $x \in \mathbb{R}_+^p$ denote a vector of p input quantities and let $y \in \mathbb{R}_+^q$, denote a vector of q output quantities. The production set consists of the set of physically (or technically) attainable combinations (x, y). For purposes of comparing the performances of producers, the technology (i.e., the efficient boundary of Ψ) given by

$$\Psi^\partial = \{(x, y) \in \Psi | (\theta x, \gamma y) \notin \Psi \text{ for all } \theta \in (0, 1) \text{ and } (x, \lambda y) \notin \Psi \text{ for all } \lambda \in (1, \infty)\} \tag{2}$$

is a relevant benchmark.[4] Firms that are technically efficient operate in the set Ψ^{∂}, while those that are technically inefficient operate in the set $\Psi \setminus \Psi^{\partial}$.[5]

The following economic assumptions are standard in microeconomic theory of the firm (e.g., Shephard 1970; Färe 1988).

Assumption 2.1 Ψ is closed.

Assumption 2.2 Both inputs and outputs are freely disposable: if $(x, y) \in \Psi$, then for any (x', y') such that $x' \geq x$ and $y' \leq y$, $(x', y') \in \Psi$.

Assumption 2.3 All production requires use of some inputs: $(x, y) \notin \Psi$ if $x = 0$ and $y \geq 0,\ y \neq 0$.

Closedness of the attainable set Ψ is a mild technical condition, avoiding mathematical problems for infinite production plans. Assumption 2.2 is sometimes called strong disposability and amounts to an assumption of monotonicity of the technology. This property also implies the technical possibility of wasting resources, i.e., the possibility of increasing input levels without producing more output, or the possibility of producing less output without reducing levels. Assumption 2.3 means that some amount of input is required to produce any output, i.e., there are no "free lunches."

The Debreu–Farrell (Debreu 1951; Farrell 1957) input measure of *technical* efficiency for a given point $(x, y) \in \mathbb{R}_+^{p+q}$ is given by

$$\theta(x, y|\Psi) = \inf\{\theta|(\theta x, y) \in \Psi\}. \tag{3}$$

Note that this measure is defined for some points in $\mathbb{R}_+^{p+q}$ not necessarily in Ψ (i.e., points for which a solution exists in (3)). Given an output level y, and an input mix (a direction) given by the vector x, the corresponding efficient level of input is given by

$$x^{\partial}(y) = \theta(x, y|\Psi)x, \tag{4}$$

which is the projection of (x, y) onto the efficient boundary Ψ^{∂}, along the ray x orthogonal to the vector y.

[4]In the literature, producers are often referred to as "decision-making units" reflecting the fact that depending on the setting, producers might be firms, government agencies, branches of firms, countries, individuals, or other agents or entities. From this point onward, we will use "firms," which are shorter than "decision-making units," to refer to producers without loss of generality or understanding.

[5]Standard microeconomic theory of the firm (e.g., Varian 1978) suggests that inefficient firms are driven out of competitive markets. However, the same theory makes no statement about how long this might take. Even in perfectly competitive markets, it is reasonable to believe that inefficient firms exist.

In general, for $(x,y) \in \Psi, \theta(x,y|\Psi)$ gives the feasible proportionate reduction of inputs that a unit located at (x, y) could undertake to become technically efficient. By construction, for all $(x,y) \in \Psi, \theta(x,y|\Psi) \in (0,1]$; (x,y) is technically efficient if and only if $\theta(x,y|\Psi) = 1$. This measure is the reciprocal of the Shephard (1970) input distance function.

Similarly, in the output direction, the Debreu–Farrell output measure of technical efficiency is given by

$$\lambda(x,y|\Psi) = \sup\{\lambda|(x,\lambda y) \in \Psi\} \tag{5}$$

for $(x,y) \in \mathbb{R}_+^{p+q}$. Analogous to the input-oriented case described above, $\lambda(x,y|\Psi)$ gives the feasible proportionate increase in outputs for a unit operating at $(x,y) \in \Psi$ that would achieve technical efficiency. By construction, for all $(x,y) \in \Psi, \lambda(x,y|\Psi) \in [1,\infty)$ and (x, y) is technically efficient if and only if $\lambda(x,y|\Psi) = 1$.

The output efficiency measure $\lambda(x,y|\Psi)$ is the reciprocal of the Shephard (1970) output distance function. The efficient level of output, for the input level x and for the direction of the output vector determined by y, is given by

$$y^{\partial}(x) = \lambda(x,y|\Psi)y. \tag{6}$$

Other distance measures have been proposed in the economic literature, including hyperbolic distance

$$\gamma(x,y|\Psi) = \sup\left\{\gamma > 0|(\gamma^{-1}x,\ \gamma y) \in \Psi\right\} \tag{7}$$

due to Färe et al. (1985) (see also Färe and Grosskopf 2004), where input and output quantities are adjusted simultaneously to reach the boundary along a hyperbolic path. Note $\gamma(x,y|\Psi) = 1$ if and only if (x, y) belongs to the efficient boundary Ψ^{∂}. Under constant returns to scale (CRS), it is straightforward to show that

$$\gamma(x,y|\Psi) = \theta(x,y|\Psi)^{-1/2} = \lambda(x,y|\Psi)^{1/2}. \tag{8}$$

However, no general relationship between the hyperbolic measure and either the input- or the output-oriented measures hold if the technology does not exhibit CRS everywhere.

Chambers et al. (1998) proposed the directional measure

$$\delta\left(x,y|\mathrm{d}_x,\mathrm{d}_y,\Psi\right) = \sup\left\{\delta|(x-\delta\mathrm{d}_x, y+\delta\mathrm{d}_y) \in \Psi\right\}, \tag{9}$$

which measures the distance from a point (x, y) to the frontier in the given direction $d = (-\mathrm{d}_x, \mathrm{d}_y)$ where $\mathrm{d}_x \in \mathbb{R}_+^p$ and $\mathrm{d}_y \in \mathbb{R}_+^q$. This measure is flexible in the sense that some values of the direction vector can be set to zero. A value $\delta(x, y|\mathrm{d}_x, \mathrm{d}_y, \Psi) = 0$ indicates an efficient point lying on the boundary of Ψ. Note that as a special case, the Debreu–Farrell radial distances can be recovered; e.g., if $d = (-x, 0)$ then $\delta(x, y|\mathrm{d}_x, \mathrm{d}_y, \Psi) = 1 - \theta(x, y|\Psi)^{-1}$ or if $d = (0, y)$ then $\delta(x, y|\mathrm{d}_x, \mathrm{d}_y, \Psi) = \lambda(x, y|\Psi) - 1$. Another interesting feature is that directional distances are additive measures; hence, they permit negative values of x and y (e.g., in finance, an output y may be the return of a fund, which can be, and often is, negative).[6] Many choices of the direction vector are possible (e.g., a common one for all firms, or a specific direction for each firm; see Färe et al. 2008 for discussion), although care should be taken to ensure that the chosen direction vector maintains invariance with respect to units of measurement for input and output quantities.

All of the efficiency measures introduced above characterize the efficient boundary by measuring distance from a known, fixed point (x, y) to the unobserved boundary Ψ^{∂}. The only difference among the measures in (3)–(9) is in the direction in which distance is measured. Consequently, the remainder of this chapter will focus on the output direction, with references given as appropriate for details on the other directions.

2.2 Statistical Model

The economic model described above introduces a number of concepts, in particular the production set Ψ and the various measures of efficiency, that are unobservable and must be estimated from data, i.e., from a sample $\mathcal{S}_n = \{(X_i, Y_i)\}_{i=1}^n$ of n-observed input–output pairs. Nonparametric estimators of the efficiency estimators introduced in Sect. 2.1 are based either on the free-disposal hull (FDH), the convex hull, or the conical hull of the sample observations. These estimators are referred to below as FDH, VRS-DEA (where "VRS" denotes variable returns to scale and "DEA" denotes data envelopment analysis) and CRS-DEA estimators, and are discussed explicitly later in Sect. 3.

[6]The measure in (9) differs from the "additive" measure $\zeta(x, y|\Psi) = \sup\{(i_{p'}\mathrm{d}_x + i_{q'}\mathrm{d}_y| (x - \mathrm{d}_x, y + \mathrm{d}_y) \in \Psi\}$ estimated by Charnes et al. (1985), where i_p, i_q denote $(p \times 1)$ and $(q \times 1)$ vectors of ones. Charnes et al. (1985) present only an estimator and do not define the object that is estimated. Moreover, the additive measure is not in general invariant to units of measurement.

Before describing the estimators, it is important to note that the theoretical results of Bahadur and Savage (1956) make clear the need for a statistical model. Such a model is defined here through the assumptions that follow, in conjunction with the assumptions discussed above. The assumptions given here correspond to those in Kneip et al. (2015b).[7] The assumptions are stronger than those used by Kneip et al. (1998, 2008), Park et al. (2000, 2010) to establish consistency, rates of convergence, and limiting distributions for FDH and DEA estimators, but are needed to establish results on moments of the estimators.

Assumption 2.4 (i) The sample observations $\mathcal{S}_n = \{(X_i, Y_i)\}_{i=1}^n$ are realizations of independent, identically distributed (iid) random variables (X, Y) with joint density f and compact support $\mathcal{D} \subset \Psi$; and (ii) f is continuously differentiable on $\mathcal{D}$.

The compact set $\mathcal{D}$ is introduced in Assumption 2.4 for technical reasons and is used in proofs of consistency of DEA and FDH estimators; essentially, the assumption rules out use of infinite quantities of one or more inputs.

Assumption 2.5 (i) $\mathcal{D}^* := \{(x, \lambda(x, y|\Psi)y)|(x, y) \in \mathcal{D}\} \subset \mathcal{D}$; (ii) $\mathcal{D}^*$ is compact; and (iii) $f(x, \lambda(x, y|\Psi)y) > 0$ for all $(x, y) \in \mathcal{D}$.

Assumption 2.5(i) ensures that the projection of any firm in $\mathcal{D}$ onto the frontier in the output direction also is contained in $\mathcal{D}$, and part (ii) means that the set of such projections is both closed and bounded. Part (iii) of the assumption ensures that the density $f(x, y)$ is positive along the frontier where firms in $\mathcal{D}$ are projected, and together with Assumption 2.4, this precludes a probability mass along the frontier. Consequently, observation of a firm with no inefficiency is an event with zero measure, i.e., any given firm almost surely operates in the interior of Ψ.

Assumption 2.6 $\lambda(x, y|\Psi)$ is three times continuously differentiable on $\mathcal{D}$.

Assumption 2.6 amounts to an assumption of smoothness of the frontier. Kneip et al. (2015b) require only two-times differentiability to establish the limiting distribution of the VRS-DEA estimator, but more smoothness is required to establish moments of the estimator. In the case of the FDH estimator, Assumption 2.6 can be replaced by the following.

[7] Kneip et al. (2015b) work in the input orientation. Here, assumptions are stated for the output orientation where appropriate or relevant.

Assumption 2.7 (i) $\lambda(x, y|\Psi)$ is twice continuously differentiable on $\mathcal{D}$; and (ii) all the first-order partial derivatives of $\lambda(x, y|\Psi)$ with respect to x and y are nonzero at any point $(x, y) \in \mathcal{D}$.

Note that the free disposability assumed in Assumption 2.2 implies that $\lambda(x, y|\Psi)$ is monotone, increasing in x and monotone, decreasing in y. Assumption 2.7 additionally requires that the frontier is strictly monotone and does not possess constant segments (which would be the case, e.g., if outputs are discrete as opposed to continuous, as in the case of ships produced by shipyards). Finally, part (i) of Assumption 2.7 is weaker than Assumption 2.6; here, the frontier is required to be smooth, but not as smooth as required by Assumption 2.6.[8]

For the VRS-DEA estimator, the following assumption is needed.

Assumption 2.8 $\mathcal{D}$ is almost strictly convex; i.e., for any $(x, y), (\tilde{x}, \tilde{y}) \in \mathcal{D}$ with $\left(\frac{x}{\|x\|}, y\right) \neq \left(\frac{\tilde{x}}{\|\tilde{x}\|}, \tilde{y}\right)$, the set $\{(x^*, y^*)|(x^*, y^*) = (x, y) + \alpha((\tilde{x}, \tilde{y}) - (x, y)) \text{ for some } 0 < \alpha < 1\}$ is a subset of the interior of $\mathcal{D}$.

Assumption 2.8 replaces the assumption of strict convexity of Ψ used in Kneip et al. (2008) to establish the limiting distribution of the VRS-DEA estimator. Together with Assumption 2.5, Assumption 2.8 ensures that the frontier Ψ^{∂} is convex in the region where observations are projected onto Ψ^{∂} by $\lambda(x, y|\Psi)$. Note, however, that convexity may be a dubious assumption in some situations, for example, see Bogetoft (1996), Bogetoft et al. (2000), Briec et al. (2004), and Apon et al. (2015). Convexity is not needed when FDH estimators are used, and in Sect. 6.1, a statistical test of the assumption is discussed.

For the case of the CRS-DEA estimator, Assumption 2.8 must be replaced by the following condition.

Assumption 2.9 (i) For any $(x, y) \in \Psi$ and any $a \in [0, \infty)$, $(ax, ay) \in \Psi$; (ii) the support $\mathcal{D} \subset \Psi$ of f is such that for any $(x, y), (\tilde{x}, \tilde{y}) \in \mathcal{D}$ with $\left(\frac{x}{\|x\|}, \frac{y}{\|y\|}\right) \neq \left(\frac{\tilde{x}}{\|\tilde{x}\|}, \frac{\tilde{y}}{\|\tilde{y}\|}\right)$, the set $\{(x^*, y^*)|(x^*, y^*) = (x, y) + \alpha((\tilde{x}, \tilde{y}) - (x, y)) \text{ forsome } 0 < \alpha < 1\}$ is a subset of the interior of $\mathcal{D}$; and (iii) $(x, y) \notin \mathcal{D}$ for any $(x, y) \in \mathbb{R}^p_+ \times \mathbb{R}^q$ with $y^1 = 0$, where y^1 denotes the first element of the vector y.

[8] Assumption 2.7 is slightly stronger, but much simpler than Assumptions AII–AIII in Park et al. (2000).

The conditions on the structure of Ψ (and $\mathcal{D}$) given in Assumptions 2.8 and 2.9 are incompatible. It is not possible that both assumptions hold simultaneously. Assumption 2.9(i) implies that the technology Ψ^{∂} is characterized by globally CRS. Under Assumption 2.9(i), Ψ is equivalent to its convex cone, denoted by $(\mathcal{V}\Psi)$. Otherwise, $\Psi \subset (\mathcal{V}\Psi)$ and (provided Assumptions 2.5 and 2.8 hold) Ψ^{∂} is said to be characterized by VRS.

Assumptions 2.4–2.6 are similar to assumptions needed by Kneip et al. (2008) to establish the limiting distribution of the VRS-DEA estimator, except that there and as noted above, the efficiency measure is only required to be twice continuously differentiable. Here, the addition of Assumption 2.8 (or Assumption 2.9 in the CRS case) and the additional smoothness of $\lambda(x, y|\Psi)$ in Assumption 2.6 are needed to establish results beyond those obtained in Kneip et al. (2008).

2.3 An Alternative Probabilistic Framework

The description of the DGP in Sect. 2.2 is traditional. However, the DGP can also be described in terms that allow a probabilistic interpretation of the Debreu–Farrell efficiency scores, providing a new way of describing the nonparametric FDH and DEA estimators. This new formulation is useful for introducing extensions of the FDH and DEA estimators described above, and for linking frontier estimation to extreme value theory as explained by Daouia et al. (2010). The presentation here draws on that of Daraio and Simar (2005), who extend the ideas of Cazals et al. (2002).

The stochastic part of the DGP introduced in Sect. 2.2 through the probability density function $f(x, y)$ is completely described by the distribution function

$$H_{XY}(x, y) = \Pr(X \leq x, Y \geq y). \tag{10}$$

This is not a standard distribution function since the cumulative form is used for the inputs x while the survival form is used for the outputs y. However, $H_{XY}(x, y)$ is well-defined and gives the probability that a unit operating at input, output levels (x, y) is *dominated,* i.e., that another unit produces at least as much output while using no more of any input than the unit operating at (x, y). The distribution function is monotone, non-decreasing in x and monotone non-increasing in y. In addition, the support of the distribution function $H_{XY}(\cdot, \cdot)$ is the attainable set Ψ, i.e.,

$$H_{XY}(x,y) = 0 \; \forall \; (x,y) \notin \Psi. \tag{11}$$

The joint probability $H_{XY}(x,y)$ can be decomposed using Bayes' rule to obtain

$$H_{XY}(x,y) = \underbrace{\Pr(X \leq x | Y \geq y)}_{=F_{X|Y}(x|y)} \underbrace{\Pr(Y \geq y)}_{=S_Y(y)} \tag{12}$$

$$= \underbrace{\Pr(Y \geq y | X \leq x)}_{=S_{Y|X}(y|x)} \underbrace{\Pr(X \leq x)}_{=F_X(x)}, \tag{13}$$

where $S_Y(y) = \Pr(Y \geq y)$ denotes the survivor function of Y, $S_{Y|X}(y|x) = \Pr(Y \geq y | X \leq x)$ denotes the conditional survivor function of Y, and the conditional distribution and survivor functions are assumed to exist whenever used (i.e., when needed, $S_Y(y) > 0$ and $F_X(x) > 0$). The frontier Ψ^{∂} can be defined in terms of the conditional distributions defined by (12) and (13) since the support of $H(x, y)$ is the attainable set. This permits definition of some alternative concepts of efficiency.

For the output-oriented case, assuming $F_X(x) > 0$, define

$$\begin{aligned} \tilde{\lambda}(x,y|H_{XY}) &= \sup\{\lambda | S_{Y|X}(\lambda y|x) > 0\} \\ &= \sup\{\lambda | H_{XY}(x, \lambda y) > 0\}. \end{aligned} \tag{14}$$

The output efficiency score $\tilde{\lambda}(x,y|H_{XY})$ gives the proportionate increase in outputs required for the same unit to have zero probability of being dominated by a randomly chosen unit, holding input levels fixed. Note that in a multivariate framework, the radial nature of the Debreu–Farrell measures is preserved. Similar definitions are possible in the input, hyperbolic and directional orientations (see Simar and Wilson (2013) for details and references).

From the properties of the distribution function $H_{XY}(x,y)$, it is clear that the new efficiency score defined in (14) has some sensible properties. First, $\tilde{\lambda}(x,y|H_{XY})$ is monotone, non-decreasing in x and monotone, non-increasing in y. Second, and most importantly, under Assumption 2.2 it is trivial to show that $\tilde{\lambda}(x,y|H_{XY}) = \lambda(x,y|\Psi)$. Therefore, whenever Assumption 2.2 holds, the probabilistic formulation presented here provides an alternative characterization of the traditional Debreu–Farrell efficiency scores.

2.4 Partial Frontiers

As will be seen in Sect. 3, estimators of the various efficiency measures introduced above are sensitive to outliers and are plagued by slow convergence rates that depend on $(p + q)$, i.e., the number of inputs and outputs. One approach to avoid these problems is to measure efficiency relative to some benchmark other than the *full frontier* Ψ^{∂}. The notion of *partial frontiers* described below provides useful, alternative benchmarks against which producers' performances can be measured. The main idea is to replace the full frontier with a model feature that lies "close" to the full frontier. Using estimates of distance to a partial frontier, one can rank firms in terms of their efficiency just as one can do when measuring distance to the full frontier.

The density of inputs and outputs introduced in Assumption 2.4 and the corresponding distribution function in (10) permit definition of other benchmarks (in addition to Ψ^{∂}) for evaluating producers' performances. Two classes of partial frontiers have been proposed: (i) order-m frontiers, where m can be viewed as a trimming parameter, and (ii) order-α quantile frontiers, analogous to traditional quantile functions but adapted to the frontier problem. As will be seen in Sects. 3.3 and 3.4, estimation based on the concept of partial frontiers offers some advantages overestimation based on the full frontier Ψ^{∂}. In particular, partial frontiers are less sensitive to outliers in the data, avoided and root-n convergence rates are achieved by estimators based on partial frontier concepts.

Suppose input levels x in the interior of the support of X are given, and consider m iid random variables Y_i, $i = 1, \ldots, m$ drawn from the conditional q-variate distribution function $F_{Y|X}(y|x) = 1 - S_{Y|X}(y|x) = \Pr(Y \leq y | X \leq x)$. Define the random set

$$\Psi_m(x) = \bigcup_{i=1}^{m} \left\{ (\tilde{x}, \tilde{y}) \in \mathbb{R}_+^{p+q} \middle| \tilde{x} \leq x, Y_i \leq \tilde{y} \right\}. \tag{15}$$

Then, the *random* set $\Psi_m(y)$ is the free-disposal hull of m randomly-chosen firms that use no more than x levels of the p inputs. For any y and the given x, the Debreu–Farrell output efficiency score relative to the set $\Psi_m(x)$ is simply

$$\lambda(x, y|\Psi_m(x)) = \sup\{\lambda | (x, \lambda y) \in \Psi_m(x)\} \tag{16}$$

after substitution of $\Psi_m(x)$ for Ψ in (5). Of course, the efficiency score $\lambda(x, y|\Psi_m(x))$ is random, since the set $\Psi_m(x)$ is random. For a given realization of the m values Y_i, a realization of $\lambda(x, y|\Psi_m(x))$ is determined by

$$\lambda(x,y|\Psi_m(x)) = \max_{i=1,\ldots,m}\left\{\min_{j=1,\ldots,p}\left(\frac{Y_i^j}{y^j}\right)\right\} \tag{17}$$

where superscript j indexes elements of the q-vectors y and Y_i. Then for any $y \in \mathbb{R}_+^q$, the (expected) order-m output efficiency measure is defined for all x in the interior of the support of X by

$$\begin{aligned}\lambda_m(x,y|H_{XY}) &= E(\lambda_m(x,y|\Psi_m(x))|X \le x)\\ &= \int_0^\infty \left[1-\left(1-S_{Y|X}(uy|x)\right)^m\right]\mathrm{d}u\\ &= \lambda(x,y|\Psi) - \int_0^{\lambda(x,y|\Psi)} \left(1-S_{Y|X}(uy|x)\right)^m;\mathrm{d}u\end{aligned} \tag{18}$$

provided the expectation exists. Clearly,

$$\lim_{m\to\infty}\lambda_m(x,y|H_{XY}(x,y)) = \lambda(x,y|\Psi). \tag{19}$$

The order-m output efficiency score provides a benchmark for the unit operating at (x, y) relative to the expected maximum output among m peers drawn randomly from the population of units that use no more than x levels of inputs. This efficiency measure in turn defines an order-m output-efficient frontier. For any $(x, y) \in \Psi$, the expected maximum level of outputs of order-m for a unit using input level x and for an output mix determined by the vector y is given by

$$y_m^\partial(x) = \lambda(x,y|\Psi_m(x))y. \tag{20}$$

The expected output-efficient frontier of order m (i.e., the output order-m efficient frontier) is defined by

$$\Psi_m^{\partial\text{out}} = \{(\tilde{x},\tilde{y})|\tilde{x} = x, \tilde{y} = y\lambda_m(x,y), (x,y) \in \Psi\}. \tag{21}$$

By contrast, recall that the full frontier Ψ^∂ is defined (at input level x) by (6).

Extension to the input-oriented case is straightforward (see Cazals et al. (2002) for details). Extension to hyperbolic and directional distances is somewhat more complicated due to the nature of the order-m concept in the multivariate framework and requires some additional work. Results are given by Wilson (2011) for the hyperbolic case, and by Simar and Vanhems (2012) for directional cases.

The central idea behind order-m estimation is to substitute a feature "close" to the frontier Ψ^∂ for the frontier itself. Introduction of the

distribution function $H_{XY}(x, y)$ invites a similar, alternative approach based on quantiles. As noted above, the quantity m in order-m frontier estimation serves as a trimming parameter which determines the percentage of points that will lie above the order-m frontier. The idea underlying order-α quantile frontiers is to reverse this causation and choose the proportion of the data lying above the partial frontier directly.

Quantile estimation in regression contexts is an old idea. In the framework of production frontiers, using the probabilistic formulation of the DGP developed in Sect. 2.3, it is straightforward to adapt the order-m ideas to order-α quantile estimation. These ideas are developed for the univariate case in the input and output orientations by Aragon et al. (2005) and extended to the multivariate setting by Daouia and Simar (2007). Wheelock and Wilson (2008) extended the ideas to the hyperbolic orientation, and Simar and Vanhems (2012) extended the ideas to directional measures.

Consider again the output-oriented case. For all x such that $F_X(x) > 0$ and for $\alpha \in (0, 1]$, the output α-quantile efficiency score for the unit operating at $(x, y) \in \Psi$ is defined by

$$\lambda_\alpha(x, y|H_{XY}) = \sup\{\lambda | S_{Y|X}(\lambda y|x) > 1 - \alpha\}. \tag{22}$$

To illustrate this measure, suppose that $\lambda_\alpha(x, y) = 1$. Then, the unit operating at $(x, y) \in \Psi$ is said to be output-efficient at the level $(\alpha \times 100)$-percent, meaning that the unit is dominated with probability $(1 - \alpha)$ by firms using *no more than* x level of inputs. More generally, if $\lambda_\alpha(x, y)(<, >)1$, the firm at (x, y) can (decrease, increase) its output to $\lambda_\alpha(x, y)y$ to become output-efficient at level $(\alpha \times 100)$-percent, i.e., to be dominated by firms using weakly less input (than the level x) with probability $(1 - \alpha)$.

The concept of order-α output efficiency allows definition of the corresponding efficient frontier at the level $(\alpha \times 100)$-percent. For a given (x, y), the order-α output efficiency level of outputs is given by

$$y_\alpha^\partial(x) = \lambda_\alpha(x, y|H_{XY})y. \tag{23}$$

By construction, a unit operating at the point $(x, y_\alpha^\partial(x)) \in \Psi$ has a probability $H_{XY}\left(x, y_\alpha^\partial(x)\right) = (1 - \alpha)F_X(x) \leq 1 - \alpha$ of being dominated. Analogous to $\Psi_m^{\partial \text{out}}$, $y_\alpha^\partial(x)$ can be evaluated for all possible x to trace out an order-α output frontier, denoted by $\Psi_\alpha^{\partial \text{out}}$.

It is straightforward to show that $\lambda_\alpha(x, y|H_{XY})$ converges monotonically to the Debreu–Farrell output efficiency measure, i.e.,

$$\lim_{\alpha \uparrow 1} \lambda_\alpha(x, y|H_{XY}) = \lambda(x, y|\Psi) \tag{24}$$

where "↑" denotes monotonic convergence from below. Moreover, for all $(x, y) \in \Psi$, $(x, y) \notin \Psi^{\partial}$, there exists an $\alpha \in (0, 1]$ such that $\lambda_{\alpha}(x, y|H_{XY}) = 1$, where $\alpha = 1 - S_{Y|X}(y|x)$.

From the preceding discussion, it is clear that both the output order-m and output order-α partial frontiers must lie weakly below the full frontier. Consequently, both $\lambda_m(x, y|H_{XY}(x, y))$ and $\lambda_{\alpha}(x, y|H_{XY}(x, y))$ must be weakly less than $\lambda(x, y|\Psi)$. Conceivably, $\lambda_m(x, y|H_{XY}(x, y))$ or $\lambda_{\alpha}(x, y|H_{XY}(x, y))$ could be less than one, indicating that (x, y) lies above the corresponding partial frontier. This is natural since by construction, the partial frontiers (provided $\alpha < 1$ or m is finite) lie in the interior of Ψ. As noted above, partial frontiers provide alternative benchmarks against which efficiency can be measured, with the advantage that the partial frontiers can be estimated with root-n convergence rate and with less sensitivity to outliers than estimates based on full frontiers.

Similar ideas have been developed for other directions. For the input orientation, see Daouia and Simar (2007). Wheelock and Wilson (2008) extend the ideas of Daouia and Simar (2007) to the hyperbolic direction, while Simar and Vanhems (2012) extend these ideas to the directional case.

It is important to note that all of the concepts discussed so far describe unobservable features of the statistical model. Consequently, efficiency must be estimated, and then inference must be made before anything can be learned about efficiency. The preceding discussion makes clear that more than one benchmark is available by which efficiency can be measured, e.g., the full frontier Ψ^{∂}, or the order-m or order-α partial frontiers. In any case, these are not observed and must be estimated. The next section discusses various nonparametric estimators of efficiency.

3 Nonparametric Estimation of Efficiency

3.1 Free-Disposal Hull Estimators

The FDH estimator

$$\hat{\Psi}_{\text{FDH}} = \bigcup_{(X_i, Y_i) \in \mathcal{S}_n} \left\{ (x, y) \in \mathbb{R}_+^{p+q} \middle| x \geq X_i, y \leq Y_i \right\}, \tag{25}$$

of Ψ proposed by Deprins et al. (1984) relies on free disposability assumed in Assumption 2.2 and does not require convexity of Ψ. The FDH estimator defined in (25) is simply the free-disposal hull of the observed sample

$\mathcal{S}_n$ and amounts to the union of n southeast-orthants with vertices (X_i, Y_i) where n is the number of input–output pairs in $\mathcal{S}_n$.

A nonparametric estimator of output efficiency for a given point $(x, y) \in \mathbb{R}_+^{p+q}$ is obtained by replacing the true production set Ψ in (5) with the estimator $\hat{\Psi}_{\text{FDH}}$, yielding

$$\hat{\lambda}_{\text{FDH}}(x, y|\mathcal{S}_n) = \sup\Big\{ \lambda | (x, \lambda y) \in \hat{\Psi}_{\text{FDH}} \Big\}. \tag{26}$$

This can be computed quickly and easily by first identifying the set

$$D(x, y) = \{i | (X_i, Y_i) \in_n,\ X_i \leq x,\ Y_i \geq y\} \tag{27}$$

of indices of observations (X_i, Y_i) that dominate (x, y) and then computing

$$\hat{\lambda}_{\text{FDH}}(x, y|\mathcal{S}_n) = \max_{i \in D(x,y)} \quad \min_{j=1,\ldots,q} \left(\frac{Y_i^j}{y^j} \right). \tag{28}$$

Determining the set $D(x, y)$ requires only some partial sorting and some simple logical comparisons. Consequently, the estimator $\hat{\lambda}_{\text{FDH}}(x, y|\mathcal{S}_n)$ can be computed quickly and easily. Efficient output levels for given input levels x and a given output mix (direction) described by the vector y are estimated by

$$\hat{y}^{\partial}(x) = \hat{\lambda}_{\text{FDH}}(x, y|\mathcal{S}_n) y. \tag{29}$$

By construction, $\hat{\Psi}_{\text{FDH}} \subseteq \Psi$, and so $\hat{\Psi}_{\text{FDH}}$ is a downward, inward-biased estimator of Ψ, and $\hat{y}^{\partial}(x)$ is a downward-biased estimator of $y^{\partial}(x)$ defined in (6). Consequently, the level of technical efficiency is under-stated by $\hat{\lambda}_{\text{FDH}}(x, y|\mathcal{S}_n)$.

The plug-in principle used above can be used to define an FDH estimator of the input-oriented efficiency measure in (3). Wilson (2011) extends this to the hyperbolic case, and Simar and Vanhems (2012) extend the idea to the directional case.

Results from Park et al. (2000) and Daouia et al. (2017) for the input-oriented case extend to the output-oriented case described here. In particular,

$$n^{1/(p+q)} \Big(\hat{\lambda}_{\text{FDH}}(x, y|\mathcal{S}_n) - \lambda(x, y|\Psi) \Big) \xrightarrow{\mathcal{L}} \text{Weibull}(\mu_{xy}, p + q) \tag{30}$$

where μ_{xy} is a constant that depends on the DGP (see Park et al. (2000) for details). Wilson (2011) and Simar and Vanhems (2012) establish similar

results for the hyperbolic and directional cases. Regardless of the direction, the various FDH efficiency estimators converge at rate $n^{1/(p+q)}$. For $(p+q) > 2$, the FDH estimators converge slower than the parametric root-n rate.

3.2 Data Envelopment Analysis Estimators

Although DEA estimators were first used by Farrell (1957) to measure technical efficiency for a set of observed firms, the idea did not gain widespread notice until Charnes et al. (1978) appeared. Charnes et al. used the convex cone (rather than the convex hull) of $\hat{\Psi}_{\text{FDH}}$ to estimate Ψ, which would be appropriate only if returns to scale are everywhere constant. Later, Banker et al. (1984) used the convex hull of $\hat{\Psi}_{\text{FDH}}$ to estimate Ψ, thus allowing variable returns to scale.[9] Here, "DEA" refers to both of these approaches, as well as other approaches that involve definition of a convex set enveloping the FDH estimator $\hat{\Psi}_{\text{FDH}}$ to estimate Ψ.

The most general DEA estimator of the attainable set Ψ is simply the convex hull of $\hat{\Psi}_{\text{FDH}}$, i.e.,

$$\hat{\Psi}_{\text{VRS}} = \left\{ (x,y) \in \mathbb{R}^{p+q} \middle| y \leq \omega,\ x \geq \omega,\ i_{n'}\omega = 1,\ \omega \in \mathbb{R}^n_+ \right\} \tag{31}$$

where $\boldsymbol{X} = [X_1 \ldots, X_n]$ and $\boldsymbol{Y} = [Y_1 \ldots, Y_n]$ are $(p \times n)$ and $(q \times n)$ matrices, respectively, whose columns are the input–output combinations in $\mathcal{S}_n$, ω is an $(n \times 1)$ vector of weights, and i_n is an $(n \times 1)$ vector of ones. Alternatively, the conical hull of the FDH estimator, $\hat{\Psi}_{\text{CRS}}$, used by Charnes et al. (1978), is obtained by dropping the constraint $i_n\omega = 1$ in (31), i.e.,

$$\hat{\Psi}_{\text{CRS}} = \left\{ (x,y) \in \mathbb{R}^{p+q} \middle| y \leq \boldsymbol{Y}\omega, x \geq \boldsymbol{X}\omega, \omega \in \mathbb{R}^n_+ \right\}. \tag{32}$$

As with the FDH estimators, DEA estimators of the efficiency scores $\theta(x,y|\Psi)$, $\lambda(x,y|\Psi)$ $\gamma(x,y|\Psi)$, and $\delta(x,y|\boldsymbol{u},\boldsymbol{v},\Psi)$ defined in (3), (5), (7), and (9) can be obtained using the plug-in method by replacing the true, but unknown, production set Ψ with one of the estimators $\hat{\Psi}_{\text{VRS}}$ or $\hat{\Psi}_{\text{CRS}}$. In the output orientation, replacing Ψ with $\hat{\Psi}_{\text{VRS}}$ in (5) yields

$$\hat{\lambda}_{\text{VRS}}(x,y|\mathcal{S}_n) = \sup\left\{ \lambda | (x,\lambda y) \in \hat{\Psi}_{\text{VRS}} \right\}, \tag{33}$$

[9]Confusingly, both Charnes et al. (1978) and Banker et al. (1984) refer to their estimators as "models" instead of "estimators." The approach of both is a-statistical, but the careful reader will remember that truth cannot be learned from data.

which can be computed by solving the linear program

$$\hat{\lambda}_{\text{VRS}}(x, y|\mathcal{S}_n) = \left\{ \lambda|\lambda y \leq \boldsymbol{Y}\omega, x \geq \boldsymbol{X}\omega, i_n'\omega = 1, \omega \in \mathbb{R}_+^n \right\}. \quad (34)$$

The CRS estimator $\hat{\lambda}_{\text{CRS}}(x, y|\mathcal{S}_n)$ is obtained by dropping the constraint $i_n'\omega = 1$ on the right-hand side (RHS) of (34). Technically efficient levels of outputs can be estimated by plugging either the VRS or CRS version of the DEA efficiency estimator into (6). For example, under VRS, in the output orientation the technically efficient level of inputs for a given level of inputs x is estimated by $\hat{\lambda}_{\text{VRS}}(x, y|\mathcal{S}_n)y$.

These ideas extend naturally to the input orientation as well as to the directional case. In both cases, the resulting estimators based on either $\hat{\Psi}_{\text{VRS}}$ or $\hat{\Psi}_{\text{CRS}}$ can be written as linear programs. In the hyperbolic case, $\hat{\gamma}_{\text{CRS}}(x, y|\mathcal{S}_n)$ can be computed as the square root of $\hat{\lambda}_{\text{CRS}}(x, y|\mathcal{S}_n)$ due to (8) in situations where Assumption 2.9(i) is maintained. Otherwise, Wilson (2011) provides a numerical method for computing the hyperbolic estimator $\hat{\gamma}_{\text{VRS}}(x, y|\mathcal{S}_n)$ based on $\hat{\Psi}_{\text{VRS}}$.

Both the FDH and DEA estimators are biased by construction since $\hat{\Psi}_{\text{FDH}} \subseteq \hat{\Psi}_{\text{VRS}} \subseteq \Psi$. Moreover, $\hat{\Psi}_{\text{VRS}} \subseteq \hat{\Psi}_{\text{CRS}}$. Under Assumption 2.9(i), $\hat{\Psi}_{\text{CRS}} \subseteq \Psi$; otherwise, $\hat{\Psi}_{\text{CRS}}$ will not be a statistically consistent estimator of Ψ. Of course, if Ψ is not convex, then $\hat{\Psi}_{\text{VRS}}$ will also be inconsistent. These relations further imply that $\hat{\lambda}_{\text{FDH}}(x, y|\mathcal{S}_n) \leq \hat{\lambda}_{\text{VRS}}(x, y|\mathcal{S}_n) \leq \lambda(x, y|\Psi)$ and $\hat{\lambda}_{\text{VRS}}(x, y|\mathcal{S}_n) \leq \hat{\lambda}_{\text{CRS}}(x, y|\mathcal{S}_n)$. Similar relations hold for estimators of the input, hyperbolic and directional efficiency measures.

Kneip et al. (1998) derive the rate of convergence of the input-oriented DEA estimator, while Kneip et al. (2008) derive its limiting distribution. These results extend to the output orientation after straightforward (though perhaps tedious) changes in notation to establish that for $(x, y) \in \Psi$, and conditions satisfied by the assumptions listed in Sect. 2,

$$n^{2/(p+q+1)}\left(\hat{\lambda}_{\text{VRS}}(x, y|\mathcal{S}_n) - (x, y|\Psi)\right) \xrightarrow{\mathcal{L}} Q_{xy,\text{VRS}}(\cdot) \quad (35)$$

as $n \to \infty$ where $Q_{xy,\text{VRS}}(\cdot)$ is a regular, non-degenerate distribution with parameters depending on the characteristics of the DGP and on (x, y). Wilson (2011) establishes similar results for the hyperbolic orientation, and Simar et al. (2012) extend these results to the directional case. The convergence rate remains the same in both cases. Under similar assumptions but with the addition of CRS in Assumption 2.9, Park et al. (2010) establish

$$n^{2/(p+q)}\left(\hat{\lambda}_{\text{CRS}}(x, y|\mathcal{S}_n) - \lambda(x, y|\Psi)\right) \xrightarrow{\mathcal{L}} Q_{xy,\text{CRS}}(\cdot) \quad (36)$$

for $(x, y) \in \Psi$ as $n \to \infty$, where $Q_{xy,\mathrm{CRS}}(\cdot)$ is another regular, non-degenerate distribution with parameters depending on the characteristics of the DGP and on (x, y) but different from $Q_{xy,1}(\cdot)$. Interestingly, results from Kneip et al. (2016) for the input-oriented case can be extended to the output orientation (again, after straightforward but tedious changes in notation) to establish that under Assumption 2.9 and appropriate regularity conditions (see Kneip et al. 2016 for details),

$$\hat{\lambda}_{\mathrm{VRS}}(x, y|\mathcal{S}_n) - \lambda(x, y|\Psi) = O_P\left(n^{-2/(p+q)}\right). \tag{37}$$

Thus, under CRS, the VRS-DEA estimator attains the faster rate of the CRS-DEA estimator. The FDH estimator, however, keeps the same convergence rate $n^{1/(p+q)}$ regardless of returns to scale or whether Ψ is convex.

Kneip et al. (2015b) provide results on the moments of both FDH and DEA estimators as discussed in Sect. 5. Limiting distributions of the FDH efficiency estimators have a Weibull form, but with parameters that are difficult to estimate. The limiting distributions of the DEA estimators do not have a closed form. Hence, in either case, inference on individual efficiency scores requires bootstrap techniques. In the DEA case, Kneip et al. (2008) provide theoretical results for both a smoothed bootstrap and for subsampling, while Kneip et al. (2011) and Simar and Wilson (2011b) provide details and methods for practical implementation. Subsampling can also be used for inference in the FDH case (see Jeong and Simar (2006) and Simar and Wilson (2011b).

3.3 Order-*m* Estimators

In Sects. 3.1 and 3.2, the plug-in approach was used to define estimators of the output efficiency measure defined in (5) based on the full frontier. A similar approach is used here to define an estimator of $\lambda_m(x, y|H_{XY})$ defined in (18). The empirical analog of the distribution function $H_{XY}(x, y)$ defined in (11) is

$$\hat{H}_{XY,n}(x, y) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x, Y_i \ge y), \tag{38}$$

where $I(\cdot)$ is the indicator function (i.e., $I(A) = 1$ if A; otherwise, $I(A) = 0$). Note that at any point $(x, y) \in \mathbb{R}^{p+q}$, $\hat{H}_{XY,n}(x, y)$ gives the proportion of sample observations in $\mathcal{S}_n$ with values $X_i \le x$ and $Y_i \ge y$; in other words,

$\hat{H}_{XY,n}(x,y)$ gives the proportion of points in the sample $\mathcal{S}_n$ that (weakly) dominate (x,y). In addition,

$$\hat{S}_{y|x,n}(y|x) = \frac{\hat{H}_{XY,n}(x,y)}{\hat{H}_{XY,n}(x,0)} \tag{39}$$

is the empirical analog of the conditional survivor function $S_{Y|X}(y|x)$ defined in (13).

Substitution of (39) into the second line of (18) yields a nonparametric estimator of $\lambda_m(x,y|H_{XY})$, namely

$$\lambda_{m,n}\left(x,y|\hat{H}_{XY,n}\right) = \int_0^\infty \left[1-\left(1-\hat{S}_{Y,n}(uy|x)\right)^m\right]\mathrm{d}u. \tag{40}$$

The integral in (40) is univariate and can be evaluated using numerical integration methods such as Gaussian quadrature. Alternatively, the estimator can be computed using Monte Carlo methods as described by Cazals et al. (2002). For a firm operating at $(x,y) \in \mathcal{P}$, an estimate of its expected maximum output level of order-m is given by

$$\hat{y}_m^{\partial}(x) = \lambda\left(x,y|\hat{H}_{XY,n}\right)y, \tag{41}$$

analogous to (20).

Cazals et al. (2002) show that for finite m, $\lambda_m\left(x,y|\hat{H}_{XY,n}\right)$ is asymptotically normal with root-n convergence rate, i.e.,

$$\sqrt{n}\left(\lambda_m(x,y|\hat{H}_{XY,n}) - \lambda_m(x,y)\right) \xrightarrow{\mathcal{L}} N\left(0,\sigma_m^2(x,y)\right) \tag{42}$$

as $n \to \infty$. An expression for the variance term σ_m^2 is given by Cazals et al. (2002). Although the convergence rate does not depend on the dimensionality (i.e., on $p+q$), the variance increases with $p+q$ when the sample size n is held fixed. It is not hard to find examples where most if not all the observations in a sample lie above the order-m frontier unless m is very large.

Cazals et al. (2002) discuss the input-oriented analog of the estimator in (40). Wilson (2011) extends these ideas to the hyperbolic case, and Simar and Vanhems (2012) extend the ideas to directional distances. In each case, the estimators achieve strong consistency with root-n rate of convergence and asymptotic normality when m is finite. In addition, as

$m \to \infty$, $\lambda_m(x,y) \to \lambda(x,y|\Psi)$ and $\lambda_{m,n}\left(x,y|\hat{H}_{XY,n}\right) \to \hat{\lambda}_{FDH}(x,y_n)$. If $m = m(n) \to \infty$ at rate $n\log nF_X(x)$ as $n \to \infty$,

$$n^{\frac{1}{p+q}}\left[\lambda_{m,n}\left(x,y|\hat{H}_{XY,n}\right) - \lambda(x,y|\Psi)\right] \xrightarrow{\mathcal{L}} \text{Weibull}(\mu_{x,y}, p+q) \text{ as } n \to \infty. \tag{43}$$

As $m = m(n) \to \infty$, the estimator $\lambda_{m,n}\left(x,y|\hat{H}_{XY,n}\right)$ shares the same properties as the FDH estimator. But, in finite samples, it will be more robust to outliers and extreme values since it will not envelop all the observations in the sample.

3.4 Order-α Estimators

Here also, the plug-in principle can be used to obtain nonparametric order-α efficiency estimators. A nonparametric estimator of the output α-quantile efficiency score defined in (22) is obtained by replacing the conditional survival function in (22) with its empirical analog introduced in Sect. 3.3, yielding

$$\lambda_\alpha\left(x,y|\hat{H}_{XY,n}\right) = \sup\left\{\lambda|\hat{S}_{Y|X,n}(\lambda y|x)(\lambda y|x) > 1-\alpha\right\}. \tag{44}$$

The order-α efficiency estimator in (44) is easy to compute using the algorithm given by Daouia and Simar (2007). Define

$$\mathcal{Y}_i = \min_{k=1,\ldots,q} \frac{Y_i^k}{y^k}, \tag{45}$$

$i = 1,\ldots,n$, and let $n_x = n\hat{F}_X(x) > 0$ where $\hat{F}_X(x)$ is the p-variate empirical distribution function of the input quantities, i.e., the empirical analog of $F_X(x)$ defined by (13). For $j = 1, \ldots, n_x$, let $\mathcal{Y}_{(j)}^x$ denote the jth order statistic of the observations $\mathcal{Y}_i$ such that $X_i \leq x$, so that $\mathcal{Y}_{(1)}^x \leq \mathcal{Y}_{(2)}^x \leq \ldots \leq \mathcal{Y}_{(n_x)}^x$. Then,

$$\hat{\lambda}_{\alpha,n} = \begin{cases} \mathcal{Y}_{(\alpha n_x)}^x \text{ if } \alpha n_x \in \mathbb{N}_{++}; \\ \mathcal{Y}_{(\lfloor \alpha n_x \rfloor + 1)}^x \text{ otherwise,} \end{cases} \tag{46}$$

where $\lfloor \alpha n_x \rfloor$ denotes the integer part of αn_x and $\mathbb{N}_{++}$ is the set of strictly positive integers.

The input-oriented estimator is obtained similarly (see Daouia and Simar (2007) for details). The ideas are extended to the hyperbolic case by Wheelock and Wilson (2008), who also present a numerical procedure for computing the estimator that is faster than the method based on sorting by a factor of about 70. The directional case is treated by Simar and Vanhems (2012).

Regardless of the direction that is chosen, provided $\alpha < 1$, the order-α estimators converge completely; e.g., in the output orientation,

$$\lambda_{\alpha,n}\left(x, y|H_{XY,n}\right) \xrightarrow{\rfloor} \lambda_{\alpha}(x, y), \tag{47}$$

where $\xrightarrow{\rfloor}$ denotes complete convergence. Complete convergence, introduced by Hsu and Robbins (1947), implies and is a stronger form of convergence than almost-sure convergence. A sequence of random variables $\{\zeta_n\}_{n=1}^{\infty}$ converges completely to a random variable ζ, denoted by $\zeta_n \xrightarrow{\rfloor} \zeta$, if $\lim_{n\to\infty} \sum_{j=1}^{n} \Pr\left(|\zeta_j - \zeta| \geq \varepsilon\right) < \infty \; \forall \; \varepsilon > 0$. The complete convergence in (47) implies

$$\lim_{n\to\infty} \sum_{j=1}^{n} \Pr\left(\left|\lambda_{\alpha,j}\left(x, y|\hat{H}_{XY,n}\right) - \lambda_{\alpha}(x, y)\right| \geq \varepsilon\right) < \infty \tag{48}$$

for all $\varepsilon > 0$. In addition, for $\alpha < 1$, the order-α estimator is asymptotically normally distributed, with root-n convergence rate:

$$\sqrt{n}\left(\lambda_{\alpha,n}\left(x, y|\hat{H}_{XY,n}\right) - \lambda_{\alpha}(x, y)\right) \xrightarrow{\mathcal{L}} N\left(0, \sigma_{\alpha}^{2}(x, y)\right), \text{ as } n \to \infty. \tag{49}$$

An expression is for the variance term $\sigma_{\alpha}^{2}(x, y)$ is given by Daouia and Simar (2007). Note, however, that similar to the order-m estimator, as $\alpha \to 1$ the order-α estimator $\lambda_{\alpha,n}\left(x, y|\hat{H}_{XY,n}\right)$ shares the properties of the FDH estimator.

3.5 Making Inference About Efficiency of a Firm

The root-n rate of convergence of the order-m and order-α estimators is unusual among nonparametric estimators. Asymptotic normality of the partial efficiency estimators facilitates inference and allows estimation of confidence intervals using an asymptotic normal approximation. In the case of the order-m estimator, Cazals et al. (2002) describe a simple plug-in method for estimating the variance parameter $\sigma_{m}^{2}(x, y)$ that appears in (42). A similar

approach can be used to estimate the variance term $\sigma_\alpha^2(x,y)$ that appears in (49). In either case, due to the asymptotic normality of the order-m and order-α estimators, the variance terms can also be estimated using a standard, naive bootstrap where observations are resampled uniformly, with replacement to create bootstrap samples of size n.

When either FDH or DEA estimators are used, making inference is more problematic. Although the limiting distributions of the FDH and DEA estimators have been derived, they are complicated. The limiting distribution of the FDH estimator is a Weibull, but its parameter depends on unknown model features (e.g., curvature of the true, unobserved frontier) that are difficult to estimate. In the DEA case, the limiting distributions do not have closed-form expressions. The only practical approach is to make inference using bootstrap methods. However, it is well known (e.g., see Bickel and Freedman 1981) that an ordinary, naive bootstrap does not provide valid inference when estimating a support boundary (or distance to the support boundary). Simar and Wilson (2011a) provide a pedagogical explanation of the problem. Unfortunately, this fact is not recognized in some of the frontier literature as indicated by the discussion in Simar and Wilson (1999a, b).

Simar and Wilson (1998) propose a smooth bootstrap to replace the inconsistent naive bootstrap when FDH or DEA estimators are used. Simar and Wilson implement the smoothed bootstrap in a simple model under the assumption that the distribution of the inefficiencies along the chosen direction (input rays or output rays) is homogeneous in the input–output space. Hence, the smoothing operates only on the estimation of the univariate density of the efficiencies, making the problem much easier to handle. Simar and Wilson (2000) extend this idea to a more general heterogeneous case where the distribution of efficiency is allowed to vary over Ψ. This requires more complication than the original procedure and involves the estimation of a smoothed density of (X, Y) with unknown support in a $(p + q)$ -dimensional space. No theoretical justification was given for either approach, but results from intensive Monte Carlo experiments described in both papers suggest that these bootstrap procedures give reasonable approximations for correcting the bias of the efficiency estimates and for building individual confidence intervals for the efficiency of any fixed point (x, y).

Kneip et al. (2008) describe two bootstrap techniques and prove that both provide consistent, valid inference. The first is a double-smooth bootstrap where in addition to smoothing the empirical distribution of the data, the support of Ψ is estimated by a smoothed version of the VRS-DEA estimator. The second is a subsampling bootstrap.

The double-smooth bootstrap developed by Kneip et al. (2008) involves numerical difficulties making it difficult to implement and computationally demanding. Kneip et al. (2011) provide a simplified, consistent, and computationally efficient version of the double-smooth bootstrap. The idea is rather simple. It is well known that the naive bootstrap does not work, but it does not work only because of points in a neighborhood of the boundary. The idea behind the simplified Kneip et al. (2011) method is to draw naively among observations which are "far" from the frontier and draw the remaining points from a uniform distribution with support "near" the frontier. This neighborhood of the frontier is tuned by a smoothing parameter that can be selected by a simple "rule of thumb." For obtaining consistency, the VRS-DEA frontier estimate must be smoothed, and here, a second bandwidth parameter is selected by cross-validation methods.

The subsampling approach is much easier to implement, since a bootstrap sample $\mathcal{S}_{\tilde{n}}^{*}$, where $\tilde{n} = n^{\gamma}$ for some $\gamma \in (0, 1)$, is obtained by drawing (either with or without replacement) $\tilde{n}$ pairs (X_i, Y_i) from the original sample $\mathcal{S}_n$. Kneip et al. (2008) prove consistency of this subsampling bootstrap, but do not provide suggestions for how a value for $\tilde{n}$ might be selected in practice. Their simulation results indicate that performance of the subsampling bootstrap in terms of achieved coverages of estimated confidence intervals is quite sensitive to the choice of $\tilde{n}$. Jeong and Simar (2006) prove that subsampling provides a consistent approximation of the sampling distribution of the FDH efficiency estimator, but here again, no practical advice is offered on how to select an appropriate subsample size.

Using results from Politis et al. (2001) and Bickel and Sakov (2008), Simar and Wilson (2011a) provide a data-driven algorithm for selecting an appropriate value of the subsample size $\tilde{n}$, for both the FDH and DEA cases. The idea is to compute the object of interest (e.g., bounds of a confidence interval, or bias estimate) for various values of $\tilde{n}$ on some selected grid. Then, the value of $\tilde{n}$ where the results show the smallest volatility is selected. This volatility can be computed for each value of $\tilde{n}$ in the grid by computing, for example, the standard deviation between the 3 or 5 values found for the adjacent values of $\tilde{n}$. Simar and Wilson (2011a) investigate the performance of their method (in terms of achieved coverages of individual confidence intervals for efficiency scores) by intensive Monte Carlo experiments, for both FDH and DEA estimators. The results indicate that the method works well for moderate sample sizes similar to those faced in practice, providing reasonable approximations of the sampling distribution of the estimators. The method has been implemented in the FEAR software package described by Wilson (2008).

4 Environmental Factors

4.1 A Statistical Model with Environmental Variables

The statistical model presented in Sect. 2 includes only input quantities and output quantities represented by random variables $X \in \mathbb{R}_+^p$ and $Y \in \mathbb{R}_+^q$, respectively. But in some situations, the production process may be affected by influences beyond inputs and outputs. For example, firms may face different regulatory environments. In the US hospital industry, hospital type, or ownership (i.e., for-profit, nonprofit, state and local government, or federal government) may influence incentives, objectives, or other features that might affect production. Similarly, in banking and insurance, private, public, or mutual ownership might influence production. In agricultural studies, rainfall might be regarded as something other than an input, exogenous to farmers' choices. Below, we refer to such factors—that are neither inputs nor outputs—as *environmental factors,* and formalize a statistical model of the production process along the lines of the probability framework of Cazals et al. (2002) and Daraio et al. (2018).

Let $Z \in \mathbb{R}^r$ denote a vector of variables describing environmental factors faced by producers, with random variables (X, Y, Z) defined on an appropriate probability space. Let $f_{XYZ}(x, y, z)$ denote the joint density of (X, Y, Z) which has support $\mathcal{P} \subset \mathbb{R}_+^p \times \mathbb{R}_+^q \times \mathbb{R}^r$. This joint density can always be decomposed as

$$f_{XYZ}(x, y, z) = f_{XY|Z}(x, y|z) f_Z(z). \tag{50}$$

Let

$$\Psi^z = \{(X, Y) | X \text{ can produce } Y \text{ when } Z = z\} \tag{51}$$

be the set of feasible combinations of inputs and outputs for a firm facing environmental conditions $Z = z$. Then, Ψ^z is the conditional support of $f_{XY|Z}(x, y|z)$, i.e., the support of (x, y) given $Z = z$. Let $\mathcal{Z}$ denote the support of $f_Z(z)$.

There are several ways in which the environmental variables in Z might affect the production process. First, the environmental variables might affect only the inefficiency process through the density $f_{XY|Z}(x, y|z)$, thereby affecting the probability for a firm to be within some neighborhood of the frontier. Second, the environmental variables might operate only through Ψ^z, thereby determining the support of (X, Y) but not the level of inefficiency. In other words, the environmental variables might affect production possibilities, but

not the level of inefficiency. A third possibility is that the variables in Z affect both the level of inefficiency as well as the support of (X, Y).

Let

$$\Psi^+ = \bigcup_{z \in \mathcal{Z}} \Psi^z. \tag{52}$$

By construction, $\Psi^z \subseteq \Psi^+ \subset \mathbb{R}_+^{p+q}$ for all $z \in \mathbb{R}$. Then, *one and only one* of the following conditions must be satisfied.

Assumption 4.1 (*Separability Condition*): $\Psi^z = \Psi^+$ for all $z \in \mathcal{Z}$.

Assumption 4.2 (*Non-Separability Assumption*): $\Psi^z \neq \Psi^+$ for some $z \in \mathcal{Z}$.

Assumptions 4.1 and 4.2 are mutually exclusive and collectively exhaustive. Assumption 4.1 is referred to as the "separability" condition by Simar and Wilson (2007).

Under Assumption 4.1, it is clear that the joint support $\mathcal{P}$ of (X, Y, Z) can be factorized as

$$\mathcal{P} = \Psi^+ \times \mathcal{Z}, \tag{53}$$

and Ψ^+ is equivalent to Ψ defined by (1).[10] However, $\Psi^+ = \Psi$ *if and only if* Assumption 4.1 holds. Alternatively, under Assumption 4.2, Ψ is not well defined since in this case whether X can produce Y depends necessarily depends upon Z. Under Assumption 4.2, Ψ^+ remains well defined but contains pairs (X, Y) that are not feasible. Moreover, under Assumption 4.2, Ψ^+ does not describe any useful feature of the model and has no useful economic interpretation.

Under Assumption 4.1, one can perform a two-stage analysis where the environmental variables in Z are ignored in the first stage and efficiency is estimated by one of the *unconditional* estimators $\hat{\lambda}_{\text{FDH}}(x, y|\mathcal{S}_n)$, $\hat{\lambda}_{\text{VRS}}(x, y|\mathcal{S}_n)$, or $\hat{\lambda}_{\text{CRS}}(x, y|\mathcal{S}_n)$. Then, the set of resulting efficiency estimates can be regressed in a second stage on the environmental variables Z using a truncated regression model as explained by Simar and Wilson (2007). Note, however, that conventional inference (e.g., using estimates of

[10] Assumption 4.1 is referred to as the "separability condition" by Simar and Wilson (2007) because of (53), i.e., since the support of (X, Y, Z) can be written as the Cartesian product of the production set $\Psi^+ = \Psi$ and the support of Z when Assumption 4.1 holds.

variance obtained by inverting the negative Hessian of the log-likelihood for the truncated regression) is invalid in the second-stage regression. Simar and Wilson (2007) propose a bootstrap method for making inference in the second stage (see Simar and Wilson (2011b) and Kneip et al. (2015b) for additional discussion regarding inference-making in the second-stage regression).

Alternatively, if efficiency is estimated in a first stage using one of the *unconditional* efficiency estimators and then the resulting efficiency estimates are regressed on elements of Z, then neither the first-stage estimates nor the second-stage estimates estimate any useful or meaningful model feature if Assumption 4.2 holds instead of Assumption 4.1. Since Assumption 4.1 is a rather restrictive assumption, care should be taken. Simar and Wilson (2007) note that Assumption 4.1 should be tested against the alternative hypothesis given by Assumption 4.2, but such a test has only recently been provided by Daraio et al. (2018). The test is discussed in Sect. 6.4.

Since Ψ is not well defined under Assumption 4.2, the efficiency measures defined in (3), (5), (7), and (9) are also not well defined. Whenever Assumption 4.2 holds, notions of *conditional* efficiency are needed. In the output orientation, the conditional measure

$$\lambda(x, y|z) = \sup\{\lambda > 0|(x, \lambda y) \in \Psi^z\} \tag{54}$$

introduced by Cazals et al. (2002) and Daraio and Simar (2005) gives a measure of distance to the appropriate, relevant boundary (i.e., the boundary of Ψ^z that is attainable by firms operating under conditions described by z).

As in Sect. 2.3, the conditional efficiency measure in (54) can be described in probabilistic terms. The conditional density $f_{XY|Z}(x, y|z)$ in (50) implies a corresponding distribution function

$$H_{XY|Z}(x, y|z) = \Pr(X \leq x,\ Y \geq y|Z = z), \tag{55}$$

giving the probability of finding a firm dominating the production unit operating at the level (x, y) and facing environmental conditions z. Then, analogous to the reasoning in Sect. 2.3, the conditional efficiency score can be written as

$$\lambda(x, y|z) = \sup\{\lambda > 0|H_{XY|Z}(x, \lambda y|z) > 0\}. \tag{56}$$

Let h denote a vector of bandwidths of length r corresponding to elements of Z and z. In order to define a statistical model that incorporates

environmental variables, consider an h-neighborhood of $z \in \mathcal{Z}$ such that $|Z - z| \leq h$ and define the *conditional* attainable set given by

$$\begin{aligned}\Psi^{z,h} &= \{(X,Y)|X \text{ can produce } Y, \text{when} |Z-z| \leq h\} \\ &= \left\{(x,y) \in \mathbb{R}_+^{p+q} \middle| H^h_{XY|Z}(x,y|z) > 0\right\} \\ &= \left\{(x,y) \in \mathbb{R}_+^{p+q} \middle| f^h_{XY|Z}(\cdot,\cdot|z) > 0\right\}\end{aligned} \tag{57}$$

where

$$H^h_{XY|Z}(x,y|z) = \Pr(X \leq x, Y \geq y | z-h \leq Z \leq z+h) \tag{58}$$

gives the probability of finding a firm dominating the production unit operating at the level (x, y) and facing environmental conditions Z in an h-neighborhood of z. The corresponding conditional density of (X, Y) given $|Z - z| \leq h$, denoted by $f^h_{XY|Z}(\cdot,\cdot|z)$, is implicitly defined by

$$H^h_{XY|Z}(x,y|z) = \int_{-\infty}^{x} \int_{y}^{\infty} f^h_{XY|Z}(u,v|Z \in [z-h, z+h]) \mathrm{d}v \mathrm{d}u. \tag{59}$$

Clearly, $\Psi^{z,h} = \bigcup_{|\tilde{z}-z| \leq h} \Psi^{\tilde{z}}$. Following Jeong et al. (2010), define for a given h

$$\begin{aligned}\lambda^h(x,y|z) &= \sup\left\{\lambda > 0 | (x, \lambda y) \in \Psi^{z,h}\right\} \\ &= \sup\left\{\lambda > 0 | H^h_{XY|Z}(x, \lambda y|z) > 0\right\}.\end{aligned} \tag{60}$$

The idea behind estimating efficiency *conditionally* on $Z = z$ is to let the bandwidth h tend to 0 as $n \to \infty$. The idea is motivated by smoothing methods similar to those used in nonparametric regression problems, but adapted to frontier estimation. Some additional, new assumptions are needed to complete the statistical model.

The next three assumptions are conditional analogs of Assumptions 2.1–2.3 appearing in Sect. 2.1.

Assumption 4.3 For all $z \in \mathcal{Z}$, Ψ^z and $\Psi^{z,h}$ are closed.

Assumption 4.4 For all $z \in \mathcal{Z}$, both inputs and outputs are strongly disposable; i.e., for any $z \in \mathcal{Z}$, $\tilde{x} \geq x$ and $0 \leq \tilde{y} \leq y$, if $(x, y) \in \Psi^z$ then

$(\tilde{x}, y) \in \Psi^z$ and $(x, \tilde{y}) \in \Psi^z$. Similarly, if $(x, y) \in \Psi^{z,h}$ then $(x, \tilde{y}) \in \Psi^{z,h}$ and $(x, \tilde{y}) \in \Psi^{z,h}$.

Assumption 4.5 For all $z \in \mathcal{Z}$, if $x = 0$ and $y \geq 0$, $y \neq 0$ then (i) $(x, y) \notin \Psi^z$ and (ii) $(x, y) \notin \Psi^{z,h}$.

Assumption 4.4 corresponds to Assumption 1F in Jeong et al. (2010) and amounts to a regularity condition on the conditional attainable sets justifying the use of the localized versions of the FDH and DEA estimators. The assumption imposes weak monotonicity on the frontier in the space of inputs and outputs for a given $z \in \mathcal{Z}$ and is standard in microeconomic theory of the firm. Assumption 4.5 is the conditional analog of Assumption 2.3 and rules out free lunches.

The next assumption concerns the regularity of the density of Z and of the conditional density of (X, Y) given $Z = z$, as a function of z in particular near the efficient boundary of Ψ^z (see Assumptions 3 and 5 in Jeong et al. 2010).

Assumption 4.6 Z has a continuous density $f_Z(\cdot)$ such that for all $z \in f_Z(z)$ is bounded away from zero. Moreover, the conditional density $f_{XY|Z}(\cdot, \cdot|z)$ is continuous in z and is strictly positive in a neighborhood of the frontier of Ψ^z.

Assumption 4.7 For all (x, y) in the support of (X, Y), $\lambda^h(x, y|z) - \lambda(x, y|z) = O(h)$ as $h \to 0$.

Assumption 4.7 amounts to an assumption of continuity of $\lambda(\cdot, \cdot|z)$ as a function of z and is analogous to Assumption 2 of Jeong et al. (2010).

The remaining assumptions impose regularity conditions on the data-generating process. The first assumption appears as Assumption 4 in Jeong et al. (2010).

Assumption 4.8 *(i) The sample observations* $\mathcal{X}_n = \{(X_i, Y_i, Z_i)\}_{i=1}^n$ *are realizations of iid random variables* (X, Y, Z) *with joint density* $f_{XYZ}(\cdot, \cdot, \cdot)$*; and (ii) the joint density* $f_{XYZ}(\cdot, \cdot, \cdot)$ *of* (X, Y, Z) *is continuous on its support.*

The next assumptions are needed to establish results for the moments of the conditional FDH and DEA estimators described in Sect. 4.2. The assumptions here are conditional analogs of Assumptions 3.1–3.4 and 3.6, respectively, in Kneip et al. (2015b). Assumption 4.9, part (iii), and Assumption 4.10, part (iii), appear as Assumption 5 in Jeong et al. (2010).

Assumption 4.9 For all $z \in \mathcal{Z}$, (i) the conditional density $f_{XY|Z}(\cdot, \cdot|z)$ of $(X, Y)|Z = z$ exists and has support $\mathcal{D}^z \subset \Psi^z$; (ii) $f_{XY|Z}(\cdot, \cdot|z)$ is continuously differentiable on $\mathcal{D}^z$; and (iii) $f^h_{XY|Z}(\cdot, \cdot|z)$ converges to $f_{XY|Z}(\cdot, \cdot|z)$ as $h \to 0$.

Assumption 4.10 (i) $\mathcal{D}^{z*} := \{(x, \lambda(x, y|z)y)|(x, y) \in \mathcal{D}^z\} \subset \mathcal{D}^z$; (ii) $\mathcal{D}^{z*}$ is compact; and (iii) $f_{XY|Z}(x, \lambda(x, y|z)y|z) > 0$ for all $(x, y) \in \mathcal{D}^z$.

Assumption 4.11 For any $z \in \mathcal{Z}$, $\lambda(x, y|z)$ is three times continuously differentiable with respect to x and y on $\mathcal{D}^z$.

Assumption 4.12 For all $z \in \mathcal{Z}$, (i) $\lambda(x, y|z)$ is twice continuously differentiable on $\mathcal{D}^z$; and (ii) all the first-order partial derivatives of $\lambda(x, y|z)$ with respect to x and y are nonzero at any point $(x, y) \in \mathcal{D}^z$.

Assumption 4.13 For any $z \in \mathcal{Z}$, $\mathcal{D}^z$ is almost strictly convex; i.e., for any $(x, y), (\tilde{x}, \tilde{y}) \in \mathcal{D}^z$ with $\left(\frac{x}{\|x\|}, y\right) \neq \left(\frac{\tilde{x}}{\|\tilde{x}\|}, \tilde{y}\right)$, the set $\{(x^*, y^*)|(x^*, y^*) = (x, y) + \alpha(\tilde{x}, \tilde{y}) \text{ for some } \alpha \in (0, 1)\}$ is a subset of the interior of $\mathcal{D}^z$.

Assumption 4.14 For any $z \in \mathcal{Z}$, (i) for any $(x, y) \in \Psi^z$ and any $a \in [0, \infty)$, $(ax, ay) \in \Psi^z$; (ii) the support $\mathcal{D}^z \subset \Psi^z$ of $f_{XY|Z}$ is such that for any $(x, y), (\tilde{x}, \tilde{y}) \in \mathcal{D}^z$ with $\left(\frac{x}{\|x\|}, \frac{y}{\|y\|}\right) \neq \left(\frac{\tilde{x}}{\|\tilde{x}\|}, \frac{\tilde{y}}{\|\tilde{y}\|}\right)$, the set $\{(x^*, y^*)|(x^*, y^*) = (x, y) + \alpha((\tilde{x}, \tilde{y}) - (x, y)) \text{ for some } 0 < \alpha < 1\}$ is a subset of the interior of $\mathcal{D}^z$; and (iii) $(x, y) \notin \mathcal{D}^z$ for any $(x, y) \in \mathbb{R}^p_+ \times \mathbb{R}^q$ with $y^1 = 0$, where y^1 denotes the first element of the vector y.

When the conditional FDH estimator is used, Assumption 4.12 is needed; when the conditional DEA estimator is used, this is replaced by the stronger Assumption 4.11.

Note that Assumptions 4.3–4.5 and 4.8–4.14 for the model with environmental variables are analogs of Assumptions 2.1–2.3 and 2.4–2.9 for the model without environmental variables. The set of Assumptions 4.8–4.14 are stronger than set of assumptions required by Jeong et al. (2010) to prove consistency and to derive the limiting distribution for conditional efficiency estimators. The stronger assumptions given here are needed by Daraio et al. (2018) to obtain results on moments of the conditional efficiency estimators as well as the central limit theorem (CLT) results discussed in Sect. 5.2. Daraio et al. (2018) do not consider the conditional version of the CRS-DEA estimator and hence do not use Assumption 4.14 that appears above. However, the results obtained by Daraio et al. (2018) for the conditional version of the VRS-DEA estimator that are outlined in Sect. 4.2 can be extended to the CRS case while replacing Assumption 4.13 with Assumption 4.14. In addition, note that under the separability condition in Assumption 4.1, the assumptions here reduce to the corresponding assumptions in Kneip et al. (2015b) due to the discussion in Sect. 2.

4.2 Nonparametric Conditional Efficiency Estimators

As in previous cases, the plug-in approach is useful for defining estimators of the conditional efficiency score given in (54) and (56).

For the conditional efficiency score $\lambda(x, y|z)$, a smoothed estimator of $H_{XY|Z}(x, y|z)$ is needed to plug into (56), which requires the vector h of bandwidths for Z. The conditional distribution function $H_{XY|Z}(x, y|z)$ can be replaced by the estimator

$$\hat{H}_{XY|Z}(x, y|z) = \frac{\sum_{i=1}^{n} (X_i \leq x, Y_i \geq y) K_h(Z_i - z)}{\sum_{i=1}^{n} K_h(Z_i - z)}, \tag{61}$$

where $K_h(\cdot) = (h_1 \ \ldots \ h_r)^{-1} K((Z_i - z)/h)$ and the division between vectors is understood to be component-wise. As explained in the literature (e.g., see Daraio and Simar 2007b), the kernel function $K(\cdot)$ must have bounded support (e.g., the Epanechnikov kernel). This provides the output-oriented, conditional FDH estimator

$$\hat{\lambda}_{\text{FDH}}(x, y|z, \mathcal{X} \ominus_n) = \max_{i \in \mathcal{I}_1(z,h)} \left(\min_{j=1, \ldots, p} \left(\frac{Y_i^j}{y^j} \right) \right), \tag{62}$$

where

$$\mathcal{I}_1(z, h) = \{i | z - h \leq Z_i \leq z + h \cap X_i \leq x\} \tag{63}$$

is the set of indices for observations with Z in an h-neighborhood of z and for which output levels X_i are weakly less than x.

Alternatively, where one is willing to assume that the conditional attainable sets are convex, Daraio and Simar (2007b) suggest the conditional VRS-DEA estimator of $\lambda(x, y|z)$ given by

$$\hat{\lambda}_{\text{VRS}}(x, y|z, \chi_n) = \max_{\lambda, \omega_1, \ldots, \omega_n} \left\{ \begin{array}{l} \lambda > 0 | \lambda y \leq \sum\limits_{i \in \chi_2(z,h)} \omega_i Y_i, x \geq \sum\limits_{i \in \chi_2(z,h)} \omega_i X_i, \\ \text{for some } \omega_i \geq 0 \text{ such that } \sum\limits_{i \in \chi_2(z,h)} \omega_i = 1 \end{array} \right\} \tag{64}$$

where

$$\mathcal{I}_2(z, h) = \{i | z - h \leq Z_i \leq z + h\} \tag{65}$$

is the set of indices for observations with Z in an h-neighborhood of z. Note that the conditional estimators in (62) and (64) are just localized version of the unconditional FDH and VRS-DEA efficiency estimators given in (28)

and (34), where the degree of localization is controlled by the bandwidths in h. The conditional version of CRS-DEA estimator $\hat{\lambda}_{\text{CRS}}(x, y|\mathcal{S}_n)$ is obtained by dropping the constraint $\sum_{i\in\mathcal{I}_2(z,h)} \omega_i = 1$ in (64) and is denoted by $\hat{\lambda}_{\text{CRS}}(x, y|z, \mathcal{X}_n)$. Bandwidths can be optimized by least-squares cross-validation (see Daraio et al. (2018) for discussion of practical aspects).

Jeong et al. (2010) show that the conditional version of the FDH and VRS-DEA efficiency estimators share properties similar to their unconditional counterparts whenever the elements of Z are continuous. The sample size n is replaced by the effective sample size used to build the estimates, which is of order $nh_1 \ \ldots \ h_r$, which we denote as n_h. To simplify the notation, and without loss of generality, we hereafter assume that all of the bandwidths $h_j = h$ are the same, so that $n_h = nh^r$. For a fixed point (x, y) in the interior of Ψ^z, as $n \to \infty$,

$$n_h^{\kappa}\left(\hat{\lambda}(x, y|z, \mathcal{X}_n) - \lambda(x, y|z)\right) \xrightarrow{\mathcal{L}} Q_{xy|z}(\cdot) \tag{66}$$

where again $Q_{xy|z}(\cdot)$ is a regular, non-degenerate limiting distribution analogous to the corresponding one for the unconditional case. The main argument in Jeong et al. (2010) relies on the property that there are enough points in a neighborhood of z, which is obtained with the additional assumption that $f_Z(z)$ is bounded away from zero at z and that if the bandwidth is going to zero, it should not go too fast (see Jeong et al. 2010, Proposition 1; if $h \to 0$, h should be of order $n^{-\eta}$ with $\eta < r^{-1}$).

The conditional efficiency scores have also their robust versions (see Cazals et al. (2002) and Daraio and Simar (2007b) for the order-m version, and Daouia and Simar (2007) for the order-α analog). Also, conditional measures have been extended to hyperbolic distances in Wheelock and Wilson (2008) and to hyperbolic distances by Simar and Vanhems (2012).

Bădin et al. (2012, 2014) suggest useful tools for analyzing the impact of Z on the production process, by exploiting the comparison between the conditional and unconditional measures. These tools (graphical and nonparametric regressions) allow one to disentangle the impact of Z on any potential shift of the frontier or potential shift of the inefficiency distributions. Daraio and Simar (2014) provide also a bootstrap test for testing the significance of environmental factors on the conditional efficiency scores. These tools have been used in macroeconomics to gauge the effect of foreign direct investment and time on "catching-up" by developing countries (see Mastromarco and Simar [2015]).

Florens et al. (2014) propose an alternative approach for estimating conditional efficiency scores that avoids explicit estimation of a nonstandard conditional distribution (e.g., $F_{X|Y,Z}(x|y, z)$). The approach is less sensitive

to the curse of the dimensionality described above. It is based on very flexible nonparametric location-scale regression models for pre-whitening the inputs and the outputs to eliminate their dependence on Z. This allows one to define "pure" inputs and outputs and hence a "pure" measure of efficiency. The method permits returning in a second stage to the original units and evaluating the conditional efficiency scores, but without explicitly estimating a conditional distribution function. The paper proposes also a bootstrap procedure for testing the validity of the location-scale hypothesis. The usefulness of the approach is illustrated using data on commercial banks to analyze the effects of banks' size and diversity of the services offered on the production process, and on the resulting efficiency distribution.

As a final remark, note that if Assumption 4.1 holds, so that the variables in Z have no effect on the frontier, then Assumptions 4.3–4.11 can be shown to be equivalent to the corresponding conditions in Assumptions 2.1–2.9. As a practical matter, whenever Assumption 4.1 holds, least-squares cross-validation will result in bandwidths large enough so that the sets of indices $\mathcal{I}_1(z,h)$ and $\mathcal{I}_2(z,h)$ includes all integers $1, 2, \ldots, n$. In this case, the variables in Z do not affect the frontier and the conditional efficiency estimators are equivalent to the corresponding unconditional estimators.

5 Central Limit Theorems for Mean Efficiency

5.1 Mean Unconditional Efficiency

CLT results are among the most fundamental, important results in statistics and econometrics (see Spanos (1999) for detailed discussion of their historical development and their importance in inference-making). CLTs are needed for making inference about population means. In the frontier context, one might want to make inference about

$$\mu_\lambda = E(\lambda(X,Y|\Psi)) = \int_{\mathcal{D}} \lambda(x,y|\Psi) f(x,y)\mathrm{d}x\mathrm{d}y. \tag{67}$$

If $\lambda(X_i,Y_i|\Psi)$ were observed for each $(X_i,Y_i) \in \mathcal{S}_n$, then μ could be estimated by the sample mean

$$\bar{\lambda}_n = \sum_{i=1}^{n} \lambda(X_i, Y_i|\Psi). \tag{68}$$

Then under mild conditions, the Lindeberg–Feller CLT establishes the limiting distribution of $\bar{\lambda}_n$, i.e.,

$$\sqrt{n}\left(\bar{\lambda}_n - \mu_\lambda\right) \xrightarrow{\mathcal{L}} N\left(0, \sigma_\lambda^2\right) \tag{69}$$

provided

$$\begin{aligned} \sigma_\lambda^2 &= \mathrm{VAR}(\lambda(X, Y|\Psi)) \\ &= \int_{\mathcal{D}} (\lambda(x, y|\Psi) - \mu_\lambda)^2 f(x, y)\,\mathrm{d}x\,\mathrm{d}y \end{aligned} \tag{70}$$

is finite. If the $\lambda(X_i, Y_i|\Psi)$ were observed for $i = 1, \ldots, n$, one could use (69) to estimate confidence intervals for μ_λ in the usual way, relying on asymptotic approximation for finite samples. But of course the $\lambda(X_i, Y_i|\Psi)$ are *not* observed and must be estimated. Kneip et al. (2015b) show that the bias of the FDH and DEA estimators makes inference about μ_λ problematic.[11]

Kneip et al. (2015b) establish results for the moments of FDH and DEA estimators under appropriate assumptions given in Sect. 2.[12] These results are summarized by writing

$$E\left(\hat{\lambda}(X_i, Y_i|\mathcal{S}_n) - \lambda(X_i, Y_i)\right) = Cn^{-\kappa} + R_{n,\kappa}, \tag{71}$$

where $R_{n,k} = o\left(n^{-\kappa}\right)$,

$$E\left(\left(\hat{\lambda}(X_i, Y_i|\mathcal{S}_n) - \lambda(X_i, Y_i)\right)^2\right) = o\left(n^{-\kappa}\right), \tag{72}$$

[11]Kneip et al. (2015b) focus on the input orientation, but all of their results extend to the output orientation after straightforward (but perhaps tedious) changes in notation. The discussion here is in terms of the output orientation.

[12]Throughout this section, assumptions common to VRS-DEA, CRS-DEA, and FDH estimators include Assumptions 2.1–2.3 and Assumptions 2.4 and 2.5. For the VRS-DEA estimator under VRS, "appropriate assumptions" include the common assumptions as well as Assumptions 2.6 and 2.8. For the CRS-DEA estimator, "appropriate assumptions" consist of the common assumptions and Assumption 2.9. For the FDH estimator, "appropriate assumptions" consist of the common assumptions and Assumption 2.7.

and

$$\left|\mathrm{COV}\left(\hat{\lambda}(X_i, Y_i|\mathcal{S}_n) - \lambda(X_i, Y_i),\ \hat{\lambda}\left(X_j, Y_j|\mathcal{S}_n\right) - \lambda(X_j, Y_j)\right)\right| = o\left(n^{-1}\right) \tag{73}$$

for all $i, j \in \{1, \ldots, n\}$, $i \neq j$. Here, we suppress the labels "VRS," "CRS," or "FDH" on $\hat{\lambda}$. The values of the constant C, the rate κ, and the remainder term $R_{n,\kappa}$ depend on which estimator is used. In particular,

- under VRS with the VRS-DEA estimator, $\kappa = \frac{2}{(p+q+1)}$ and $R_{n,\kappa} = O(n^{-3\kappa/2}(\log n)^{\alpha_1})$, where $\alpha_1 = (p+q+4)/(p+q+1)$;
- under CRS with either the VRS-DEA or CRS-DEA estimator, $\kappa = \frac{2}{(p+q)}$ and $R_{n,\kappa} = O(n^{-3\kappa/2}(\log n)^{\alpha_2})$ where $\alpha_2 = (p+q+3)/(p+q)$; and
- under only the free disposability assumption (but not necessarily CRS or convexity) with the FDH estimator, $\kappa = \frac{1}{(p+q)}$ and $R_{n,\kappa} = O(n^{-2\kappa}(\log n)^{\alpha_3})$, where $\alpha_3 = (p+q+2)/(p+q)$.

Note that in each case, $R_{n,\kappa} = o\left(n^{-\kappa}\right)$.

The result in (73) is somewhat surprising. It is well known that FDH and DEA estimates are correlated, due to the fact that typically the estimated efficiency of several, perhaps many observations depends on a small number of observations lying on the estimated frontier; i.e., perturbing an observation lying on the estimated frontier is likely to affect estimated efficiency for other observations. The result in (73), however, indicates that this effect is negligible.

The $Cn^{-\kappa}$ term in (71) reflects the bias of the nonparametric efficiency estimators, and its interplay with the $o\left(n^{-\kappa}\right)$ expression in (72) creates problems for inference. Let

$$\hat{\mu}_n = n^{-1}\sum_{i=1}^{n}\hat{\lambda}(X_i, Y|\mathcal{S}_n). \tag{74}$$

Theorem 4.1 of Kneip et al. (2015b) establishes that $\hat{\mu}_n$ is a consistent estimator of μ under the appropriate set of assumptions, but has bias of order $Cn^{-\kappa}$. The theorem also establishes that

$$\sqrt{n}\left(\hat{\mu}_n - \mu_\lambda - Cn^{-\kappa} - R_{n,\kappa}\right) \xrightarrow{\mathcal{L}} N(0, \sigma_\lambda^2). \tag{75}$$

If $\kappa > 1/2$, the bias term in (75) is dominated by the factor $\sqrt{n}$ and thus can be ignored; in this case, standard, conventional methods based on the Lindeberg-Feller CLT can be used to estimate confidence intervals for μ_λ. Otherwise, the bias is constant if $\kappa = 1/2$ or explodes if $\kappa < 1/2$. Note

that $\kappa > 1/2$ if and only if $p + q \leq 2$ in the VRS case, or if and only if $p + q \leq 3$ in the CRS case. In the FDH case, this occurs only in the univariate case with $p = 1$, $q = 0$ or $p = 0$, $q = 1$. Replacing the scale factor $\sqrt{n}$ in (75) with n^{γ}, with $\gamma < \kappa \leq 1/2$, is not a viable option. Although doing so would make the bias disappear as $n \to \infty$, it would also cause the variance to converge to zero whenever $\kappa \leq 1/2$, making inference using the result in (75) impossible.

It is well known that the nonparametric DEA and FDH estimators suffer from the curse of dimensionality, meaning that convergence rates become slower as $(p + q)$ increases. For purposes of estimating mean efficiency, the results of Kneip et al. (2015b) indicate the curse is even worse than before, with the "explosion" of bias coming at much smaller numbers of dimensions than found in many applied studies.

In general, whenever $\kappa \leq 1/2$, the results of Kneip et al. (2015b) make clear that conventional CLTs cannot be used to make inference about the mean μ_{λ}. The problem of controlling both bias and variance, for general number of dimensions $(p + q)$, can be addressed by using a different estimator of the population mean μ_{λ} and in addition rescaling the *estimator* of μ_{λ} by an appropriate factor different from $\sqrt{n}$ when $\kappa \leq 1/2$. Consider the factor $n_{\kappa} = \lfloor n^{2\kappa} \rfloor \leq n$, where $\lfloor \alpha \rfloor$ denotes the integer part of a (note that this covers the limiting case of $\kappa = 1/2$). Then, assume that the observations in the sample $\mathcal{S}_n$ are randomly ordered and consider the latent estimator

$$\bar{\lambda}_{n_{\kappa}} = n_{\kappa}^{-1} \sum_{i=1}^{n_{\kappa}} \lambda(X_i, Y_i). \tag{76}$$

Of course, $\bar{\lambda}_{n_{\kappa}}$ is unobserved, but it can be estimated by

$$\hat{\mu}_{n_{\kappa}} = n_{\kappa}^{-1} \sum_{i=1}^{n_{\kappa}} \hat{\lambda}(X_i, Y_i | \mathcal{S}_n), \tag{77}$$

where the notation $\hat{\lambda}(X_i, Y_i | \mathcal{S}_n)$ serves to remind the reader that the individual efficiency estimates are computed from the full sample of n observations, while the sample mean is over $n_{\kappa} \leq n$ such estimates. Here again, one can use either the VRS, CRS, or FDH version of the estimator.

Theorem 4.2 of Kneip et al. (2015b) establishes that when $\kappa \leq 1/2$,

$$n^{\kappa} \left(\hat{\mu}_{n_{\kappa}} - \mu_{\lambda} - Cn^{-\kappa} - R_{n,\kappa} \right) \xrightarrow{\mathcal{L}} N(0, \sigma_{\lambda}^2) \tag{78}$$

as $n \to \infty$. Since $\sqrt{n_{\kappa}} \left(\hat{\mu}_{n_{\kappa}} - \mu_{\lambda} \right)$ has a limiting distribution with unknown mean due to the bias term $Cn^{-\kappa}$, bootstrap approaches could be used to

estimate the bias and hence to estimate confidence intervals for μ_λ. The variance could also be estimated by the same bootstrap, or by the consistent estimator

$$\hat{\sigma}^2_{\lambda,n} = n^{-1} \sum_{i=1}^{n} \left(\hat{\lambda}(X_i, Y_i | \mathcal{S}_n) - \hat{\mu}_n \right)^2. \tag{79}$$

Subsampling along the lines of Simar and Wilson (2011a) could also be used to make consistent inference about μ_λ. However, the estimator in (77) uses only a subset of the original n observations; unless n is extraordinarily large, taking subsamples among a subset of n_κ observations will likely leave too little information to provide useful inference.

Alternatively, Kneip et al. (2015b) demonstrate that the bias term $Cn^{-\kappa}$ can be estimated using a generalized jackknife estimator (e.g., see Gray and Schucany 1972, Definition 2.1). Assume again that the observations (X_i, Y_i) are randomly ordered. Let $\mathcal{S}^{(1)}_{n/2}$ denote the set of the first $n/2$ observations in $\mathcal{S}_n$, and let $\mathcal{S}^{(2)}_{n/2}$ denote the set of remaining observations from $\mathcal{S}_n$. Let

$$\hat{\mu}^*_{n/2} = \left(\hat{\mu}^{(1)}_{n/2} + \hat{\mu}^{(2)}_{n/2} \right) / 2. \tag{80}$$

Kneip et al. (2015b) show that

$$\begin{aligned} \tilde{B}_{\kappa,n} &= (2^\kappa - 1)^{-1} \left(\hat{\mu}^*_{n/2} - \hat{\mu}_n \right) \\ &= Cn^{-\kappa} + R_{n,\kappa} + o_p\left(n^{-1/2} \right) \end{aligned} \tag{81}$$

provides an estimator of the bias term $Cn^{-\kappa}$.

Of course, for n even there are $\binom{n}{n/2}$ possible splits of the sample $\mathcal{S}_n$. As noted by Kneip et al. (2016), the variation in $\tilde{B}_{\kappa,n}$ can be reduced by repeating the above steps $\kappa \ll \binom{n}{n/2}$ times, shuffling the observations before each split of $\mathcal{S}_n$, and then averaging the bias estimates. This yields a generalized jackknife estimate

$$\hat{B}_{\kappa,n} = K^{-1} \sum_{k=1}^{K} \tilde{B}_{\kappa,n,k}, \tag{82}$$

where $\tilde{B}_{\kappa,n,k}$ represents the value computed from (81) using the k th sample split.

Theorem 4.3 of Kneip et al. (2015b) establishes that under appropriate assumptions, for $\kappa \geq 2/5$ for the VRS and CRS cases or $\kappa \geq 1/3$ for the FDH case,

$$\sqrt{n}\left(\hat{\mu}_n - \hat{B}_{\kappa,n} - \mu_\lambda + R_{n,\kappa}\right) N(0, \sigma_\lambda^2). \tag{83}$$

as $n \to \infty$.

It is important to note that (83) is not valid for κ smaller than the bounds given in the theorem. This is due to the fact that for a particular definition of $R_{n,\kappa}$ (i.e., in either the VRS/CRS or FDH cases), values of κ smaller than the boundary value cause the remainder term, multiplied by $\sqrt{n}$, to diverge toward infinity. Interestingly, the normal approximation in (83) can be used with either the VRS-DEA or CRS-DEA estimators under the assumption of CRS if and only if $p + q \leq 5$; with the DEA-VRS estimator under convexity (but not CRS) if and only if $p + q \leq 4$; and with the FDH estimator assuming only free disposability (but not necessarily convexity nor CRS) if and only if $p + q \leq 3$. For these cases, an asymptotically correct $(1 - \alpha)$ confidence interval for μ_λ is given by

$$\left[\hat{\mu}_n - \hat{B}_{\kappa,n} \pm z_{1-\alpha/2}\hat{\sigma}_{\lambda,n}/\sqrt{n}\right], \tag{84}$$

where $z_{1-\alpha/2}$ is the corresponding quantile of the standard normal distribution and $\hat{\sigma}_{\lambda,n}$ is given by (79).

In cases where κ is smaller than the bounds required by (83), the idea of estimating μ_λ by a sample mean of n_κ efficiency estimates as discussed above can be used with the bias estimate in 15, leading to Theorem 4.4 of Kneip et al. (2015b), i.e., under appropriate assumptions,

$$n^\kappa \left(\hat{\mu}_{n_\kappa} - \hat{B}_{\kappa,n} - \mu_\lambda + R_{n,\kappa}\right) \xrightarrow{\mathcal{L}} N\left(0, \sigma_\lambda^2\right). \tag{85}$$

as $n \to \infty$.

Equation 85 permits construction of consistent confidence intervals for μ_λ by replacing the unknown σ_λ^2 by its consistent estimator $\hat{\sigma}_{\lambda,n}^2$. An asymptotically correct $1 - \alpha$ confidence interval for μ_λ is given by

$$\left[\hat{\mu}_{n_\kappa} - \hat{B}_{\kappa,n} \pm z_{1-\alpha/2}\hat{\sigma}_{\lambda,n}/n^\kappa\right], \tag{86}$$

where $z_{1-\alpha/2}$ is the corresponding quantile of the standard normal distribution. Here, the normal approximation can be used directly; bootstrap methods are not necessary.

Note that when $\kappa < 1/2$, the center of the confidence interval in (86) is determined by a random choice of $n_\kappa = n^{2\kappa} < n$ elements $\hat{\lambda}(X_i, Y_i|\mathcal{S}_n)$. This may be seen as arbitrary, but any confidence interval for μ_λ may be seen arbitrary in practice since asymmetric confidence intervals can be constructed by using different quantiles to establish the endpoints. The main point, however, is always to achieve a high level of coverage without making the confidence interval too wide to be informative.

Again for $\kappa < 1/2$, the arbitrariness of choosing a particular subsample of size n_κ in (86) can be eliminated by averaging the center of the interval in (86) over all possible draws (without replacement) of subsamples of size n_κ. Of course, this yields an interval centered on $\hat{\mu}_n$, i.e.,

$$\left[\hat{\mu}_n - \hat{B}_{\kappa,n} \pm z_{1-\alpha/2}\hat{\sigma}_{\lambda,n}/n^\kappa\right]. \tag{87}$$

The only difference between the intervals (86) and (87) is the centering value. Both intervals are equally informative, because they possess exactly the same length ($2z_{1-\alpha/2}\hat{\sigma}_{\lambda,n}/n^\kappa$). The interval (87) should be more accurate (i.e., should have higher coverage) because $\hat{\mu}_n$ is a better estimator of μ_λ (i.e., has less mean-square error) than $\hat{\mu}_{n_\kappa}$. If $\kappa < 1/2$, then $n_\kappa < n$, and hence, the interval in (87) contains the true value μ_λ with probability greater than $1 - \alpha$, since by the results above, it is clear that the coverage of the interval in (87) converges to 1 as $n \to \infty$. This is confirmed by the Monte Carlo evidence presented by Kneip et al. (2015b).

In cases with sufficiently small dimensions, either (83) or (85) can be used to provide different asymptotically valid confidence intervals for μ_λ. For the VRS-DEA and CRS-DEA estimators, this is possible whenever $\kappa = 2/5$ and so $n_\kappa < n$. The interval (84) uses the scaling $\sqrt{n}$ and neglects, in (83), a term $\sqrt{n}R_{n,\kappa} = O(n^{-1/10})$, whereas the interval (86) uses the scaling n^κ, neglecting in (85) a term $n^\kappa R_{n,\kappa} = O(n^{-1/5})$. We thus may expect a better approximation by using the interval (86). The same is true for the FDH case when $\kappa = 1/3$, where the interval (84) neglects terms of order $O(n^{-1/6})$ whereas the error when using (86) is only of order $O(n^{-1/3})$. These remarks are also confirmed by the Monte Carlo evidence reported by Kneip et al. (2015b).

5.2 Mean Conditional Efficiency

Daraio et al. (2018) extend the results of Kneip et al. (2015b) presented in Sect. 5 to means of conditional efficiencies. Extension to the conditional case is complicated by the presence of the bandwidth h, which impacts convergence rates. Comparing (30) and (35) with the result in (66), it is apparent that the bandwidths in the conditional estimators reduce convergence rates from n^{κ} for the unconditional estimators to $n_h^{\kappa} = n^{\kappa/(\kappa r+1)}$ for the conditional estimators. Moreover, Theorem 4.1 of Daraio et al. (2018) establishes that under Assumptions 4.3–4.6, and in addition Assumption 4.12 for the FDH case or Assumptions 4.11 and 4.13 (referred to as "appropriate conditions" throughout the remainder of this subsection), as $n \to \infty$,

$$E\left(\hat{\lambda}(X_i, Y_i|Z_i, \mathcal{X}_n) - \lambda^h(X_i, Y_i|Z_i)\right) = C_c n_h^{-\kappa} + R_{c,n_h,\kappa} \tag{88}$$

where $R_{c,n_h,\kappa} = o\left(n_h^{-\kappa}\right)$,

$$E\left(\left(\hat{\lambda}(X_i, Y_i|Z_i, \mathcal{X}_n) - \lambda^h(X_i, Y_i|Z_i)\right)^2\right) = o(n_h^{-\kappa}), \tag{89}$$

and

$$\left|\text{COV}\left(\hat{\lambda}(X_i, Y_i|Z_i, \mathcal{X}_n) - \lambda^h(X_i, Y_i|Z_i), \hat{\lambda}\left(X_j, Y_j|Z_j, \mathcal{X}_n\right) - \lambda^h\left(X_j, Y_j|Z_j\right)\right)\right| = o(n_h^{-1}) \tag{90}$$

for all $i, j \in \{1, \ldots, n\}$, $i \neq j$. In addition, for the conditional VRS-DEA estimator, $R_{c,n_h,\kappa} = O(n_h^{-3\kappa/2}(\log n_h)^{\alpha_1})$ while for the conditional FDH estimator $R_{c,n_h,\kappa} = O(n_h^{-2\kappa}(\log n_h)^{\alpha_2})$. Note that incorporation of bandwidths in the conditional estimators reduces the order of the bias from Cn^{κ} to $C_c n_h^{\kappa}$.

Let

$$\hat{\mu}_n = n^{-1} \sum_{i=1}^{n} \hat{\lambda}(X_i, Y_i|\mathcal{X}_n) \tag{91}$$

and

$$\hat{\mu}_{c,n} = n^{-1} \sum_{i=1}^{n} \hat{\lambda}(X_i, Y_i|Z_i, \mathcal{X}_n) \tag{92}$$

denote sample means of unconditional and conditional efficiency estimators. The efficiency estimators in (91) and (92) could be either FDH or VRS-DEA estimators; differences between the two are noted below when relevant. Next, define

$$\mu_c^h = E\Big(\lambda^h(X, Y|Z)\Big) = \int \lambda^h(x, y|z) f_{XYZ}(x, y, z)\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{93}$$

and

$$\sigma_c^{2,h} = \mathrm{VAR}\Big(\lambda^h(X, Y|Z)\Big) = \int_P \Big(\lambda^h(x, y|z) - \mu_c^h\Big)^2 f_{XYZ}(x, y, z)\mathrm{d}x\mathrm{d}y\mathrm{d}z, \tag{94}$$

where $\mathcal{P}$ is defined just before (50). These are the localized analogs of μ and σ^2. Next, let $\bar{\mu}_{c,n} = n^{-1}\sum_{i=1}^n \lambda^h(X_i, Y|Z_i)$. Although $\bar{\mu}_{c,n}$ is not observed, by the Lindeberg–Feller CLT

$$\frac{\sqrt{n}\big(\bar{\mu}_{c,n} - \mu_c^h\big)}{\sqrt{\sigma_c^{2,h}}} \xrightarrow{\mathcal{L}} N(0, 1) \tag{95}$$

under mild conditions.

Daraio et al. (2018) show that there can be no CLT for means of conditional efficiency estimators analogous to the result in (75) or Theorem 4.1 of Kneip et al. (2015b). There are no cases where standard CLTs with rate $n^{1/2}$ can be used with means of conditional efficiency estimators, unless Z is irrelevant with respect to the support of $(X,\ Y)$, i.e., unless Assumption 4.1 holds. Given a sample of size n, there can be no CLT for means of conditional efficiency estimators based on a sample mean of all of the n conditional efficiency estimates.

Daraio et al. (2018) consider a random subsample $\mathcal{X}_{n_h}^*$ from $\mathcal{X}_n$ of size n_h where for simplicity we use the optimal rates for the bandwidths so that $n_h = O(n^{1/(\kappa r+1)})$. Define

$$\bar{\mu}_{c,n_h} = \frac{1}{n_h} \sum_{(X_i,Y_i,Z_i)\in\mathcal{X}_{n_h}^*} \lambda^h(X_i, Y|Z_i), \tag{96}$$

$$\hat{\mu}_{c,n_h} = \frac{1}{n_h} \sum_{(X_i,Y_i,Z_i)\in\mathcal{X}_{n_h}^*} \hat{\lambda}(X_i, Y|Z_i, \mathcal{X}_n) \tag{97}$$

and let $\hat{\mu}_{c,n_h} = E\left(\hat{\mu}_{c,n_h}\right)$. Note that the estimators on the RHS of (97) are computed relative to the full sample $\mathcal{X}_n$, but the summation is over elements of the subsample $\mathcal{X}_{n_h}^*$.

Theorem 4.2 of Daraio et al. (2018) establishes that

$$\sqrt{n_h}\left(\hat{\mu}_{c,n_h} - \mu_c^h - C_c n_h^{-\kappa} - R_{c,n_h,\kappa}\right)/\sqrt{\sigma_c^{2,h}} \xrightarrow{\mathcal{L}} N(0,1), \tag{98}$$

and in addition establishes that $\sqrt{\sigma_c^{2,h}}$ is consistently estimated by

$$\hat{\sigma}_{c,n}^{2,h} = n^{-1}\sum_{i=1}^{n}\left[\hat{\lambda}(X_i, Y_i||Z_i, \mathcal{X}_n) - \hat{\mu}_{c,n}\right]^2. \tag{99}$$

The result in (98) provides a CLT for means of conditional efficiency estimators, but the convergence rate is $\sqrt{n_h}$ as opposed to $\sqrt{n}$, and the result is of practical use only if $\kappa > 1/2$. If $\kappa = 1/2$, the bias term $C_c n_h^{-\kappa}$ does not vanish. If $\kappa < 1/2$, the bias term explodes as $n \to \infty$.

Similar to Kneip et al. (2015b), Daraio et al. (2018) propose a generalized jackknife estimator of the bias term $C_c n_n^{-\kappa}$. Assume the observations in $\mathcal{X}_n$ are randomly ordered. Let $\mathcal{X}_{n/2}^{(1)}$ denote the set of the first $n/2$ observations from $\mathcal{X}_n$, and let $\mathcal{X}_{n/2}^{(2)}$ denote the set of remaining $n/2$ observations from $\mathcal{X}_n$. Note that if n is odd, $\mathcal{X}_{n/2}^{(1)}$ can contain the first $\lfloor n/2 \rfloor$ observations and $\mathcal{X}_{n/2}^{(2)}$ can contain remaining $n - \lfloor n/2 \rfloor$ observations from $\mathcal{X}_n$. The fact that $\mathcal{X}_{n/2}^{(2)}$ contains one more observation than $\mathcal{X}_{n/2}^{(1)}$ makes no difference asymptotically. Next, for $j \in \{1,\ 2\}$ define

$$\hat{\mu}_{c,n/2}^{j} = (n/2)^{-1} \sum_{(X_i,Y_i,Z_i)\in\mathcal{X}_{n/2}^{(j)}} \hat{\lambda}\left(X_i, Y_i|Z_i, \mathcal{X}_{n/2}^{(j)}\right). \tag{100}$$

Now define

$$\hat{\mu}_{c,n/2}^{*} = \left(\hat{\mu}_{c,n/2}^{1} + \hat{\mu}_{c,n/2}^{2}\right)/2. \tag{101}$$

Then,

$$\tilde{B}_{\kappa,n_h}^{c} = (2^{\kappa} - 1)^{-1}\left(\hat{\mu}_{c,n/2}^{*} - \hat{\mu}_{c,n}\right) \tag{102}$$

is an estimator of the leading bias term $C_c n_h^{-\kappa}$ in (98). Averaging over $K \ll \binom{n}{n/2}$ splits of the sample $\mathcal{X}_n$ as before yields the generalized jackknife estimator

$$\hat{B}^c_{\kappa,n_h} = K^{-1} \sum_{k=1}^{K} \tilde{B}^c_{\kappa,n_h,k}, \tag{103}$$

where $\tilde{B}^c_{\kappa,n_h,k}$ represents the value computed from (102) using the kth sample split.

Theorem 4.3 of Daraio et al. (2018) establishes that under appropriate conditions, with $\kappa = 1/(p+q) \geq 1/3$ in the FDH case or $\kappa = 2/(p+q+1) \geq 2/5$ in the VRS-DEA case,

$$\frac{\sqrt{n_h}\left(\hat{\mu}_{c,n_h} - \mu^h_c - \hat{B}^c_{\kappa,n_h} - R^*_{c,n_h,\kappa}\right)}{\sqrt{\sigma^{2,h}_c}} \xrightarrow{\mathcal{L}} N(0,1) \tag{104}$$

as $n \to \infty$. Alternatively, Theorem 4.4 of Daraio et al. (2018) establishes that under appropriate conditions, whenever $\kappa < 1/2$,

$$\frac{\sqrt{n_{h,\kappa}}\left(\hat{\mu}_{c,n_{h,\kappa}} - \mu^h_c - \hat{B}^c_{\kappa,n_h} - R^*_{c,n_h,\kappa}\right)}{\sqrt{\sigma^{2,h}_c}} \xrightarrow{\mathcal{L}} N(0,1), \tag{105}$$

as $n \to \infty$. The results in (104) and (105) provide CLTs for means of conditional efficiencies covering all values of κ and can be used to estimate confidence intervals or for testing Assumption 4.1 versus Assumption 4.2 as described in Sect. 6.4.

6 Hypothesis Testing

6.1 Testing Convexity Versus Non-Convexity

In situations where one might want to test convexity of Ψ versus non-convexity, typically a single iid sample $\mathcal{S}_n = \{(X_i, Y_i\}_{i=1}^n$ of n input–output pairs is available. Under the null hypothesis of convexity, both the FDH and VRS-DEA estimators are consistent, but under the alternative, only the FDH estimator is consistent. It might be tempting to compute the sample means

$$\hat{\mu}^{\text{full}}_{\text{VRS},n} = n^{-1} \sum_{i=1}^{n} \hat{\lambda}_{\text{VRS}}(X_i, Y | \mathcal{S}_n) \tag{106}$$

and

$$\hat{\mu}^{\text{full}}_{\text{FDH},n} = n^{-1} \sum_{(X_i,Y_i)\in n} \hat{\lambda}_{\text{FDH}}(X_i, Y_i|\mathcal{S}_n) \tag{107}$$

using the full set of observations in $\mathcal{S}_n$ and use this with (106) to construct a test statistic based on the difference $\hat{\mu}^{\text{full}}_{\text{FDH},n} - \hat{\mu}^{\text{full}}_{\text{VRS},n}$. By construction, $\hat{\lambda}_{\text{VRS}}(X_i, Y_i|\mathcal{S}_n) \geq \hat{\lambda}_{\text{FDH}}(X_i, Y_i|\mathcal{S}_n) \geq 1$ and therefore $\hat{\mu}^{\text{full}}_{\text{VRS},n} - \hat{\mu}^{\text{full}}_{\text{FDH},n} \geq 0$. Under the null, $\hat{\mu}^{\text{full}}_{\text{FDH},n} - \hat{\mu}^{\text{full}}_{\text{VRS},n}$ is expected to be "small," while under the alternative the difference is expected to be "large."

Unfortunately, such an approach is doomed to failure. Using the output-oriented analog of Theorem 4.1 of Kneip et al. (2015b) and reasoning similar to that of Kneip et al. (2016, Section 3.2), it can be shown that $n^a\left(\hat{\lambda}_{\text{VRS}}(X_i, Y_i|\mathcal{S}_n) - \hat{\lambda}_{\text{FDH}}(X_i, Y_i|\mathcal{S}_n)\right)$ converges under the null to a degenerate distribution for any power $a \leq 1/2$ of n; that is, the asymptotic variance of the statistic is zero, and the density of the statistic converges to a Dirac delta function at zero under the null, rendering inference impossible.

Kneip et al. (2016) solve this problem by randomly splitting the sample $\mathcal{S}_n$ into two mutually exclusive, collectively exhaustive parts $\mathcal{S}_{1,n_1}$ and $\mathcal{S}_{2,n_2}$ such that $\mathcal{S}_{1,n_1} \cap \mathcal{S}_{2,n_2} = \emptyset$ and $\mathcal{S}_{1,n_1} \cup \mathcal{S}_{2,n_2} = \mathcal{S}_n$. Recall that the FDH estimator converges at rate $n^{1/(p+q)}$, while the VRS-DEA estimator converges at rate $n^{2/(p+q+1)}$ under VRS or at rate $n^{2/(p+q)}$ under CRS. Kneip et al. (2016) suggest exploiting this difference by setting $n_1^{2/(p+q+1)} = \beta n_2^{1/(p+q)}$ and $n_1 + n_2 = n$ for a given sample size n, where β is a constant, and then solving for n_1 and n_2. There is no closed-form solution, but it is easy to find a numerical solution by writing $n - n_1 - \beta^{-1} n_1^{2(p+q)/(p+q+1)} = 0$; the root of this equation is bounded between 0 and $n/2$ and can be found by simple bisection. Letting n_1 equal the integer part of the solution and setting $n_2 = n - n_1$ gives the desired subsample sizes with $n_2 > n_1$. Using the larger subsample $\mathcal{S}_{2,n_2}$ to compute the FDH estimates and the smaller subsample $\mathcal{S}_{1,n_1}$ to compute the VRS-DEA estimates allocates observations from the original sample $\mathcal{S}_n$ efficiently in the sense that more observations are used to mitigate the slower convergence rate of the FDH estimator. Simulation results provided by Kneip et al. (2016) suggest that the choice of value for β matters less as sample size n increases, and that setting $\beta = 1$ gives reasonable results across various values of n and $(p + q)$.

Once the original sample has been split, the sample means

$$\hat{\mu}_{\text{VRS},n_1} = n_1^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}_{1,n_1}} \hat{\lambda}_{\text{VRS}}\left(X_i, Y_i|\mathcal{S}_{1,n_1}\right) \tag{108}$$

and

$$\hat{\mu}_{\mathrm{FDH},n_2} = n_2^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}_{2,n_2}} \hat{\lambda}_{\mathrm{FDH}}\left(X_i, Y_i|\mathcal{S}_{2,n_2}\right) \tag{109}$$

can be computed. In addition, the sample variances

$$\hat{\sigma}^2_{\mathrm{VRS},n_2} = n_2^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}_{2,n_2}} \left[\hat{\lambda}_{\mathrm{VRS}}\left(X_i, Y_i|\mathcal{S}_{2,n_2}\right) - \hat{\mu}_{\mathrm{VRS},n_2}\right]^2. \tag{110}$$

and

$$\hat{\sigma}^2_{\mathrm{FDH},n_2} = n_2^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}_{2,n_2}} \left[\hat{\lambda}_{\mathrm{FDH}}\left(X_i, Y_i|\mathcal{S}_{2,n_2}\right)(i|2,n_2) - \hat{\mu}_{\mathrm{FDH},n_2}\right]^2. \tag{111}$$

can also be computed. Now let $\kappa_1 = 2/(p+q+1)$ and $\kappa_2 = 1/(p+q)$. Then, the corresponding bias estimates $\hat{B}_{\mathrm{VRS},\kappa_1,n_1}$ and $\hat{B}_{\mathrm{FDH},\kappa_2,n_2}$ can also be computed from the two parts of the full sample, which requires further splitting both of the two parts $\mathcal{S}_{1,n_1}$ and $\mathcal{S}_{2,n_2}$ along the lines discussed in Sect. 5.1.

The rate of the FDH estimator is slower than the rate of the VRS-DEA estimator, and hence, the rate of the FDH estimator dominates. Kneip et al. (2016) show that for $(p+q) \leq 3$, the test statistic

$$\hat{\tau}_{1,n} = \frac{\left(\hat{\mu}_{\mathrm{VRS},n_1} - \hat{\mu}_{\mathrm{FDH},n_2}\right) - \left(\hat{B}_{\mathrm{VRS},\kappa_1,n_1} - \hat{B}_{\mathrm{FDH},\kappa_2,n_2}\right)}{\sqrt{\frac{\hat{\sigma}^2_{\mathrm{VRS},n_1}}{n_1} + \frac{\hat{\sigma}^2_{\mathrm{FDH},n_2}}{n_2}}} \xrightarrow{\mathcal{L}} N(0,1) \tag{112}$$

can be used to test the null hypothesis of convexity for Ψ versus the alternative hypothesis that Ψ is not convex.

Alternatively, if $(p+q) > 3$, the sample means must be computed using subsets of $\mathcal{S}_{1,n_1}$ and $\mathcal{S}_{2,n_2}$. For $\ell \in \{1,2\}$, let $\kappa = \kappa_2 = 1/(p+q)$ and for $\ell \in \{1,\ 2\}$ let $n_{\ell,\kappa} = \left\lfloor n_\ell^{2\kappa} \right\rfloor$ so that $n_{\ell,\kappa} < n_\ell$ for $\kappa < 1/2$. Let $_{\ell,n_{\ell,\kappa}}$ be a random subset of $n_{\ell,\kappa}$ input–output pairs from $\mathcal{S}_{\ell,n_\ell}$. Then, let

$$\hat{\mu}_{\mathrm{VRS},n_{1,\kappa}} = n_{1,\kappa}^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}^*_{1,n_{1,\kappa}}} \hat{\lambda}_{\mathrm{VRS}}\left(X_i, Y_i|\mathcal{S}_{1,n_1}\right) \tag{113}$$

and

$$\hat{\mu}_{\mathrm{FDH},n_{2,\kappa}} = n_{2,\kappa}^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}^*_{2,n_{2,\kappa}}} \hat{\lambda}_{\mathrm{FDH}}\left(X_i, Y_i|\mathcal{S}_{2,n_2}\right), \tag{114}$$

noting that the summations in (113) and (114) are over subsets of the observations used to compute the efficiency estimates under the summation signs.

Then by Theorem 4.4 of Kneip et al. (2015b), for $(p + q) > 3$,

$$\hat{\tau}_{2,n} = \frac{\left(\hat{\mu}_{\mathrm{VRS},n_{1,\kappa}} - \hat{\mu}_{\mathrm{FDH},n_{2,\kappa}}\right) - \left(\hat{B}_{\mathrm{VRS},\kappa_1,n_1} - \hat{B}_{\mathrm{FDH},\kappa_2,n_2}\right)}{\sqrt{\frac{\hat{\sigma}^2_{\mathrm{VRS},n_1}}{n_{1,\kappa}} + \frac{\hat{\sigma}^2_{\mathrm{FDH},n_2}}{n_{2,\kappa}}}} \xrightarrow{\mathcal{L}} N(0,1) \tag{115}$$

under the null hypothesis of convexity for Ψ. Note that $\hat{\tau}_{2,n}$ differs from $\hat{\tau}_{1,n}$ both in terms of the number of efficiency estimates used in the sample means as well as the divisors of the variance terms under the square-root sign.

Depending on whether $(p + q) \leq 3$ or $(p + q) > 3$, either $\hat{\tau}_{1,n}$ or $\hat{\tau}_{2,n}$ can be used to test the null hypothesis of CRS, with critical values obtained from the standard normal distribution. In particular, for $j \in \{1, 2\}$, the null hypothesis of convexity of Ψ is rejected if $\hat{p} = 1 - \Phi\left(\hat{\tau}_{j,n}\right)$ is less than a suitably small value, e.g., 0.1, 0.05, or 0.01.

6.2 Testing Constant Versus Variable Returns to Scale

Kneip et al. (2016) use the CLT results of Kneip et al. (2015b) discussed in Sect. 5.1 to develop a test of CRS versus VRS for the technology Ψ^{∂}. A key result is that under CRS, the VRS-DEA estimator attains same convergence rate $n^{2/(p+q)}$ as the CRS estimator as established by Theorem 3.1 of Kneip et al. (2016).

Under the null hypothesis of CRS, both the VRS-DEA and CRS-DEA estimators of $\lambda(X, Y)$ are consistent, but under the alternative, only the VRS-DEA estimator is consistent. Consider the sample means given by (106) and

$$\hat{\mu}^{\mathrm{full}}_{\mathrm{CRS},n} = n^{-1} \sum_{i=1}^{n} \hat{\lambda}_{\mathrm{CRS}}(X_i, Y_i|\mathcal{S}_n) \tag{116}$$

computed using all of the n observations in $\mathcal{S}_n$. Under the null, one would expect $\hat{\mu}^{\mathrm{full}}_{\mathrm{CRS},n} - \hat{\mu}^{\mathrm{full}}_{\mathrm{VRS},n}$ to be "small," while under the alternative

$\hat{\mu}^{\text{full}}_{\text{CRS},n} - \hat{\mu}^{\text{full}}_{\text{VRS},n}$ is expected to be "large." However, as shown by Kneip et al. (2016), a test statistic using the difference in the sample means given by (106)–(116) will have a degenerate distribution under the null since the asymptotic variance of $\hat{\mu}^{\text{full}}_{\text{CRS},n} - \hat{\mu}^{\text{full}}_{\text{VRS},n}$ is zero, similar to the case of testing convexity versus non-convexity discussed in Sect. 6.1. Consequently, subsampling methods are needed here just as they were in Sect. 6.1.

In order to obtain non-degenerate test statistics, randomly split the sample into two samples $\mathcal{S}_{1,n_1}$, $\mathcal{S}_{2,n_2}$ such that $\mathcal{S}_{1,n_1} \cup \mathcal{S}_{2,n_2} = \mathcal{S}_n$ and $\mathcal{S}_{1,n_1} \cap \mathcal{S}_{2,n_2} = \emptyset$, where $n_1 = \lfloor n/2 \rfloor$ and $n_2 = n - n_1$. Next, let

$$\hat{\mu}_{\text{CRS},n_2} = n_2^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}_{2,n_2}} \hat{\lambda}_{\text{CRS}}\left(X_i, Y_i|\mathcal{S}_{2,n_2}\right). \tag{117}$$

and

$$\hat{\sigma}^2_{\text{CRS},n_2} = n_2^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}_{2,n_2}} \left[\hat{\lambda}_{\text{CRS}}\left(X_i, Y_i|\mathcal{S}_{2,n_2}\right) - \hat{\mu}_{\text{CRS},n_2}\right]^2, \tag{118}$$

and recall (108) and (110) from the discussion in Sect. 6.1.

The (partial) sample $\mathcal{S}_{2,n_2}$ can be used to compute an estimate $\hat{B}_{\text{CRS},\kappa,n_2}$ of bias for the CRS estimator by splitting $\mathcal{S}_{2,n_2}$ into two parts along the lines discussed above. Then under the null hypothesis of CRS, Kneip et al. (2016) demonstrate that

$$\hat{\tau}_{3,n} = \frac{\left(\hat{\mu}_{\text{CRS},n_2} - \hat{\mu}_{\text{VRS},n_1}\right) - \left(\hat{B}_{\text{CRS},\kappa,n_2} - \hat{B}_{\text{VRS},\kappa,n_1}\right)}{\sqrt{\frac{\hat{\sigma}^2_{\text{VRS},n_1}}{n_1} + \frac{\hat{\sigma}^2_{\text{CRS},n_2}}{n_2}}} \xrightarrow{\mathcal{L}} N(0,1) \tag{119}$$

provided $p + q \leq 5$.

Alternatively, if $p + q > 5$, the sample means must be computed using subsets of the available observations. For $\ell \in \{1, 2\}$ and $\mathcal{S}^*_{\ell,n_{\ell,\kappa}}$ defined as in Sect. 6.1, let

$$\hat{\mu}_{\text{CRS},n_{2,\kappa}} = n^{-1}_{2,\kappa} \sum_{(X_i,Y_i)\in\mathcal{S}^*_{2,n_{2,\kappa}}} \hat{\lambda}_{\text{CRS}}\left(X_i, Y_i|_{2,n_2}\right). \tag{120}$$

and recall the definition of $\hat{\mu}_{\text{VRS},n_{1,\kappa}}$ in (113). Here again, the summation in (120) is over a random subset of the observations used to compute the efficiency estimates under the summation sign. Again under the null hypothesis

of CRS, by Theorem 4.4 of Kneip et al. (2015b), and Theorem 3.1 of Kneip et al. (2016),

$$\hat{\tau}_{4,n} = \frac{\left(\hat{\mu}_{\text{CRS},n_{2,\kappa}} - \hat{\mu}_{\text{VRS},n_{1,\kappa}}\right) - \left(\hat{B}_{\text{CRS},\kappa,n_2} - \hat{B}_{\text{VRS},\kappa,n_1}\right)}{\sqrt{\frac{\hat{\sigma}^2_{\text{VRS},n_1}}{n_{1,\kappa}} + \frac{\hat{\sigma}^2_{\text{CRS},n_2}}{n_{2,\kappa}}}} \xrightarrow{\mathcal{L}} N(0,1) \tag{121}$$

for $(p + q) > 5$. Similar to the comparison between $\hat{\tau}_{2,n}$ and $\hat{\tau}_{1,n}$ in Sect. 6.1, $\hat{\tau}_{4,n}$ differs from $\hat{\tau}_{3,n}$ both in terms of the number of efficiency estimates used to compute the sample means in the numerator as well as the divisors of the variance terms under the square-root sign in the denominator.

Depending on the value of $(p + q)$, either $\hat{\tau}_{3,n}$ or $\hat{\tau}_{4,n}$ can be used to test the null hypothesis of CRS, with critical values obtained from the standard normal distribution. In particular, for $j \in \{3, 4\}$, the null hypothesis of CRS is rejected if $\hat{p} = 1 - \Phi(\hat{\tau}_{j,n})$ is less than, say, 0.1, 0.05, or 0.01.

6.3 Testing for Differences in Mean Efficiency Across Groups of Producers

Testing for differences in mean efficiency across two groups was suggested—but not implemented—in the application of Charnes et al. (1981), who considered two groups of schools, one receiving a treatment effect and the other not receiving the treatment. There are many situations where such a test might be useful. For example, one might test whether mean efficiency among for-profit producers is greater than mean efficiency of nonprofit producers in studies of hospitals, banks and credit unions, or perhaps other industries. One might similarly be interested in comparing average performance of publicly traded versus privately held firms, or in regional differences that might reflect variation in state-level regulation or other industry features. However, the discussion in Sect. 5 makes clear that standard CLT results are not useful in general when considering means of nonparametric efficiency estimates.

Consider two groups of firms G_1 and G_2 of sizes n_1 and n_2. Suppose the researcher wishes to test whether $\mu_{1,\lambda} = E(\lambda(X,Y)|(X,Y) \in G_1)$ and $\mu_{2,\lambda} = E(\lambda(X,Y)|(X,Y) \in G_2)$ are equal against the alternative that group 1 has greater mean efficiency. More formally, one might test the null hypothesis $H_0 : \mu_{1,\lambda} = \mu_{2,\lambda}$ versus the alternative hypothesis $H_1 : \mu_{1,\lambda} > \mu_{2,\lambda}$. One could also conduct a test with a two-sided

alternative; such a test would of course follow a procedure similar to the one outlined here for a one-sided test.

The test discussed here makes no restriction on whether firms in the two groups face the same frontier, i.e., whether they operate in the same production set. Kneip et al. (2016, Section 3.1.2) discuss a version of the test where under the null firms in the two groups face the same frontier. Readers can consult Kneip et al. (2016) for details on the alternative form of the test.

Suppose iid samples $\mathcal{S}_{1,n_1} = \{(X_i, Y_i)\}_{i=1}^{n_1}$ and $\mathcal{S}_{2,n_2} = \{(X_i, Y_i)\}_{i=1}^{n_2}$ of input–output pairs from groups 1 and 2, respectively are available. In addition, assume these samples are independent of each other. The test here is simpler than the tests of convexity versus non-convexity and constant versus variable returns to scale since two independent samples are available initially. The two samples yield independent estimators

$$\hat{\mu}_{1,n_1} = n_1^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}_{1,n_1}} \hat{\lambda}\big(X_i, Y_i|\mathcal{S}_{1,n_1}\big) \tag{122}$$

and

$$\hat{\mu}_{2,n_2} = n_2^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}_{2,n_2}} \hat{\lambda}\big(X_i, Y_i|\mathcal{S}_{2,n_2}\big) \tag{123}$$

of $\mu_{1,\lambda}$ and $\mu_{2,\lambda}$, respectively; the conditioning indicates the sample used to compute the efficiency estimates under the summation signs. As in Sects. 6.1 and 6.2, the subscripts on $\hat{\lambda}(\cdot)$ have been dropped; either the FDH, VRS-DEA, or CRS-DEA estimators with corresponding convergence rates n^{κ} could be used, although the same estimator would be used for both groups. Theorem 4.1 of Kneip et al. (2015b) establishes (under appropriate regularity conditions; see Sect. 5.1 for details) consistency and other properties of these estimators. The same theorem, however, makes clear that standard, conventional central limit theorems can be used to make inference about the population means $\mu_{1,\lambda}$ and $\mu_{2,\lambda}$ only when the dimensionality $(p + q)$ is small enough so that $\kappa > 1/2$ due to the bias of the estimators $\hat{\mu}_{1,n_1}$ and $\hat{\mu}_{2,n_2}$.

As discussed earlier in Sect. 5, the Lindeberg-Feller and other central limit theorems fail when FDH, VRS-DEA, or CRS-DEA estimators are averaged as in (122)–(123) due to the fact that while averaging drives the variance to zero, it does not diminish the bias. From Kneip et al. (2015b) and the discussion in Sect. 5.1, it can be seen that unless $(p + q)$ is very small, scaling sample means such as (122)–(123) by a power of the sample size to stabilize the bias results in a degenerate statistic with the variance converging to zero. On the other hand, scaling if the power of the sample size is chosen to

stabilize the variance, the bias explodes. Hence, the bias must be estimated and dealt with.

For each of the two groups $\ell \in \{1, 2\}$, a bias estimate $\hat{B}_{\ell,\kappa,n_\ell}$ can be obtained as described earlier in Sect. 5.1. Then, using VRS-DEA estimators with $p+q \leq 4$ (or CRS-DEA estimators with $p+q \leq 5$, or FDH estimators with $p+q \leq 3$), the test statistic

$$\hat{\tau}_{5,n_1,n_2} = \frac{\left(\hat{\mu}_{1,n_1} - \hat{\mu}_{2,n_2}\right) - \left(\hat{B}_{1,\kappa,n_1} - \hat{B}_{2,\kappa,n_2}\right) - \left(\mu_{1,\lambda} - \mu_{2,\lambda}\right)}{\sqrt{\frac{\hat{\sigma}^2_{1,n_1}}{n_1} + \frac{\hat{\sigma}^2_{2,n_2}}{n_2}}} \xrightarrow{\mathcal{L}} N(0,1), \tag{124}$$

can be used to test the null hypothesis of equivalent mean efficiency for groups 1 and 2 provided $n_1/n_2 \to c > 0$ as $n_1, n_2 \to \infty$, where c is a constant. The variance estimates appearing in the denominator of $\hat{\tau}_{5,n_1,n_2}$ are given by

$$\hat{\sigma}^2_{\ell,n_\ell} = n_\ell^{-1} \sum_{i=1}^{n_\ell} \left[\hat{\lambda}\left(X_{\ell,i}, Y_{\ell,i} \middle| \mathcal{S}_{\ell,n_\ell}\right) - \hat{\mu}_{\ell,n_\ell}\right]^2 \xrightarrow{p} \sigma^2_\ell \tag{125}$$

for $\ell \in \{1,2\}$; and all values of $(p+q)$.

In situations where $p+q > 4$ with VRS-DEA estimators (or $p+q > 5$ with CRS-DEA estimators, or $p+q > 3$ with FDH estimators), means based on the subsamples must be used. For $\ell \in \{1,2\}$, let $n_{\ell,\kappa} = \left\lfloor n_\ell^{2\kappa} \right\rfloor$; then $n_{\ell,\kappa} < n_\ell$ for $\kappa < 1/2$. Let $\mathcal{S}^*_{\ell,n_{\ell,\kappa}}$ be a random subset of $n_{\ell,\kappa}$ input–output pairs from $\mathcal{S}_{\ell,n_\ell}$. Then, let

$$\hat{\mu}_{\ell,n_{\ell,\kappa}} = n_{\ell,\kappa}^{-1} \sum_{(X_{\ell,i},Y_{\ell,i}) \in \mathcal{S}^*_{\ell,n_{\ell,\kappa}}} \hat{\lambda}\left(X_{\ell,i}, Y_{\ell,i} \middle| \mathcal{S}_{\ell,n_\ell}\right), \tag{126}$$

noting that while the summation is over only the input–output pairs in $\mathcal{S}^*_{\ell,n_{\ell,\kappa}}$, the efficiency estimates under the summation sign are computed using all of the input–output pairs in $\mathcal{S}_{\ell,n_\ell}$.

Kneip et al. (2016) obtain a similar test statistic for use in situations where $p+q > 4$ with VRS-DEA estimators, $p+q > 5$ with CRS-DEA estimators, or $p+q > 3$ with FDH estimators. In particular,

$$\hat{\tau}_{6,n_{1,\kappa},n_{2,\kappa}} = \frac{\left(\hat{\mu}_{1,n_{1,\kappa}} - \hat{\mu}_{2,n_{2,\kappa}}\right) - \left(\hat{B}_{1,\kappa,n_1} - \hat{B}_{2,\kappa,n_2}\right) - \left(\mu_{1,\lambda} - \mu_{2,\lambda}\right)}{\sqrt{\frac{\hat{\sigma}^2_{1,n_1}}{n_{1,\kappa}} + \frac{\hat{\sigma}^2_{2,n_2}}{n_{2,\kappa}}}} \xrightarrow{\mathcal{L}} N(0,1), \tag{127}$$

again provided $n_1/n_2 \to c > 0$ as $n_1, n_2 \to \infty$. Note that the same estimates for the variances and biases are used in (127) as in (124). The only difference between (124) and (127) is in the number of observations used to compute the sample means and in the quantities that divide the variance terms under the square-root sign in the denominator.

Kneip et al. (2016) note that the results in (124) and (127) hold for *any* values of $\mu_{1,\lambda}$ and $\mu_{2,\lambda}$. Hence, if one tests $H_0 : \mu_{1,\lambda} = \mu_{2,\lambda}$ versus an alternative hypothesis such as $H_1 : \mu_{1,\lambda} > \mu_{2,\lambda}$ or perhaps $H_1 : \mu_{1,\lambda} \neq \mu_{2,\lambda}$, the (asymptotic) distribution of the test statistic will be known under the null and up to $\mu_{1,\lambda}$, $\mu_{2,\lambda}$ under the alternative hypothesis. Consequently, given two independent samples, one can either (i) compute under the null (so that $(\mu_{1\lambda} - \mu_{2,\lambda}) = 0$) $\hat{\tau}_{5,n_1,n_2}$ or $\hat{\tau}_{6,n_{1,\kappa},n_{2,\kappa}}$ as appropriate and compare the resulting value against a critical value from the $N(0,1)$ distribution, or (ii) use (124) and (127) to estimate a confidence interval for $(\mu_{1\lambda} - \mu_{2,\lambda})$. If the estimated interval does not include 0, one would reject the null; otherwise, one would fail to reject the null. Furthermore, the outcome will be the same under either approach; that is, for a given test size, both approaches will either reject or fail to reject the null. It clearly follows that for a given departure from the null, the tests will reject the null with probability tending to one as $n_1, n_2 \to \infty$, and hence, the tests are consistent.

6.4 Testing Separability

Daraio et al. (2018) develop a test of separability (Assumption 4.1) versus the alternative of non-separability (Assumption 4.2) using the CLT results for both unconditional and conditional efficiencies discussed in Sect. 5. The idea for building a test statistic is to compare the conditional and unconditional efficiency scores using relevant statistics that are functions of $\hat{\lambda}(X_i, Y_i|\mathcal{S}_n)$ and $\hat{\lambda}(X_i, Y_i|Z_i, \mathcal{S}_n)$ for $i = 1, \ldots, n$. Note that under Assumption 4.1, $\lambda(X, Y) = \lambda(X, Y|Z)$ with probability one, even if Z may influence the distribution of the inefficiencies inside the attainable set, and the two estimators converge to the same object. But under Assumption 4.2, the conditional attainable sets Ψ^z are different and the two estimators converge to different objects. Moreover, under Assumption 4.2, $\lambda(X, Y) \geq \lambda(X, Y|Z)$ with strict inequality holding for some $(X, Y, Z) \in \mathcal{P}$.

In order to implement the test of separability, randomly split the sample $\mathcal{S}_n$ into two independent, parts $\mathcal{S}_{1,n_1}$, $\mathcal{S}_{2,n_2}$ such that $n_1 = \lfloor n/2 \rfloor$, $n_2 = n - n_1$, $\mathcal{S}_{1,n_1} \bigcup \mathcal{S}_{2,n_2} = \mathcal{S}_n$, and $\mathcal{S}_{1,n_1} \bigcap \mathcal{S}_{2,n_2} = \emptyset$. The n_1

observations in $\mathcal{S}_{1,n_1}$ are used for the unconditional estimates, while the n_2 observations in $\mathcal{S}_{2,n_2}$ are used for the conditional estimates.[13]

After splitting the sample, compute the estimators

$$\hat{\mu}_{n_1} = n_1^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}_{1,n_1}} \hat{\lambda}\big(X_i, Y_i|\mathcal{S}_{1,n_1}\big) \tag{128}$$

and

$$\hat{\mu}_{c,n_{2,h}} = n_{2,h}^{-1} \sum_{(X_i,Y_i,Z_i)\in\mathcal{S}^*_{2,n_{2,h}}} \hat{\lambda}\big(X_i, Y_i|Z_i, \mathcal{S}_{2,n_2}\big), \tag{129}$$

where as above in Sect. 5, $\mathcal{S}^*_{2,n_{2,h}}$ in (129) is a random subsample from $\mathcal{S}_{2,n_2}$ of size $n_{2,h} = min(n_2, n_2 h^r)$. Consistent estimators of the variances are given in the two independent samples by

$$\hat{\sigma}^2_{n_1} = n_1^{-1} \sum_{(X_i,Y_i)\in\mathcal{S}_{1,n_1}} \Big(\hat{\lambda}\big(X_i, Y_i|\mathcal{S}_{1,n_1}\big)^2 - \hat{\mu}_{n_1}\Big) \tag{130}$$

and

$$\widehat{\sigma}^{2,h}_{c,n_2} = n_2^{-1} \sum_{(X_i,Y_i,Z_i)\in\mathcal{S}_{2,n_2}} \Big(\widehat{\lambda}(X_i, Y_i|Z_i, \mathcal{S}_{2,n_2}) - \widehat{\mu}_{c,n_2}\Big)^2 \tag{131}$$

respectively, where the full (sub)samples are used to estimate the variances.

The estimators of bias for a single split of each subsample for the unconditional and conditional cases are given by $\hat{B}_{\kappa,n_1}$ for the unconditional case and $\hat{B}^c_{\kappa,n_{2,h}}$ for the conditional case as described in Sects. 5.1 and 5.2, respectively.

[13] Kneip et al. (2016) proposed splitting the sample unevenly to account for the difference in the convergence rates between the (unconditional) DEA and FDH estimators used in their convexity test, giving more observations to the subsample used to compute FDH estimates than to the subsample used to compute DEA estimates. Recall that the unconditional efficiency estimators converge at rate n^κ, while the conditional efficiency estimators converge at rate n_h^κ. The optimal bandwidths are of order $n^{-\kappa/(r\kappa+1)}$, giving a rate of $n^{\kappa/(r\kappa+1)}$ for the conditional efficiency estimators. Using the logic of Kneip et al. (2016), the full sample $\mathcal{S}_n$ can be split so that the estimators in the two subsamples achieve the same rate of convergence by setting $n_1^\kappa = n_2^{\kappa/(r\kappa+1)}$. This gives $n_1 = n_2^{1/(r\kappa+1)}$. Values of n_1, n_2 are obtained by finding the root η_0 in $n - \eta - \eta^{1/(r\kappa+1)} = 0$ and setting $n_2 = \lfloor \eta_0 \rfloor$ and $n_1 = n - n_2$. However, this will often result in too few observations in the first subsample to obtain meaningful results. For example, if $p = q = r = 1$ and $n = 200$, following the reasoning above would lead to $n_1 = 22$ and $n_2 = 178$.

For values of $(p+q)$ such that $\kappa \geq 1/3$ in the FDH case or $\kappa \geq 2/5$ when DEA estimators are used, the CLT results in (83) and (104) can be used to construct an asymptotically normal test statistic for testing the null hypothesis of separability. Since the bias-corrected sample means are independent due to splitting the original sample into independent parts, and since two sequences of independent variables each with normal limiting distributions have a joint bivariate normal limiting distribution with independent marginals, if follows that for the values of $(p+q)$ given above

$$\tau_{7,n} = \frac{\left(\hat{\mu}_{n_1} - \hat{\mu}_{c,n_{2,h}}\right) - \left(\hat{B}_{\kappa,n_1} - \hat{B}^{c}_{\kappa,n_{2,h}}\right)}{\sqrt{\frac{\hat{\sigma}^2_{n_1}}{n_1} + \frac{\hat{\sigma}^{2,h}_{c,n_2}}{n_{2,h}}}} \xrightarrow{\mathcal{L}} N(0,1) \tag{132}$$

under the null. Alternatively, for $\kappa < 1/2$, similar reasoning and using the CLT results in (85) and (105) leads to

$$\tau_{8,n} = \frac{\left(\widehat{\mu}_{n_{1,\kappa}} - \widehat{\mu}_{c,n_{2,h,\kappa}}\right) - \left(\widehat{B}_{\kappa,n_1} - \widehat{B}^{c}_{\kappa,n_{2,h}}\right)}{\sqrt{\frac{\widehat{\sigma}^2_{n_1}}{n_{1,\kappa}} + \frac{\widehat{\sigma}^{2,h}_{c,n_2}}{n_{2,h,\kappa}}}} \xrightarrow{\mathcal{L}} N(0,1) \tag{133}$$

under the null, where $n_{1,\kappa} = \left\lfloor n_1^{2\kappa} \right\rfloor$ with $\hat{\mu}_{n_{1,\kappa}} = n^{-1}_{1,\kappa} \sum_{(X_i,Y_i)\in\mathcal{S}^*_{n_{1,\kappa}}} \hat{\lambda}\left(X_i, Y_i | \mathcal{S}_{n_1}\right)$, and $\mathcal{S}^*_{n_{1,\kappa}}$ is a random subsample of size $n_{1,\kappa}$ taken from $\mathcal{S}_{n_1}$ (see Kneip et al. 2015b for details). For the conditional part, we have similarly and as described in the preceding section, $n_{2,h,\kappa} = \left\lfloor n^{2\kappa}_{2,h} \right\rfloor$, with $\hat{\mu}_{c,n_{2,h,\kappa}} = n^{-1}_{2,h,\kappa} \sum_{(X_i,Y_i,Z_i)\in\mathcal{S}^*_{n_{2,h,\kappa}}} \hat{\lambda}\left(X_i, Y_i | Z_i, \mathcal{S}_{n_2}\right)$ where $\mathcal{S}^*_{n_{2,h,\kappa}}$ is a random subsample of size $n_{2,h,\kappa}$ from $\mathcal{S}_{n_2}$.

Given a random sample $\mathcal{S}_n$, one can compute values $\hat{\tau}_{7,n}$ or $\hat{\tau}_{8,n}$ depending on the value of $(p+q)$. The null should be rejected whenever $1 - \Phi\left(\hat{\tau}_{7,n}\right)$ or $1 - \Phi\left(\hat{\tau}_{8,n}\right)$ is less than the desired test size, e.g., 0.1, 0.05, or 0.01, where $\Phi(\cdot)$ denotes the standard normal distribution function.

6.5 Computational Considerations

Each of the tests described in Sects. 6.1, 6.2, and 6.4 requires randomly splitting the original sample of size n into two parts in order to compute sample means that are independent of each other. In addition, these tests

as well as the test in Sect. 6.3 require splitting the two parts randomly in order to obtain bias estimates. At several points in the preceding discussion, observations are assumed to be randomly ordered. In practice, however, data are often not randomly ordered when they are first obtained. Data may be sorted by firms' size or by some other criteria, perhaps one not represented by a variable or variables included in the researcher's data.

Observations can be randomly ordered by applying the modified Fisher-Yates shuffle (Fisher and Yates 1948) described by Durstenfeld (1964). The generalized jackknife estimates of bias involve randomly splitting groups of observations many times and then averaging, and the random splitting can be accomplished by shuffling the observations before each split. One should expect little difference in the final bias estimates between two researchers who initialize their random number generators with different seeds. However, the tests of convexity, CRS, and separability require an initial split of a single sample into two parts. Conceivably, the tests may provide different results depending on how the sample is split, making it difficult to replicate results. To avoid this problem, Daraio et al. (2018) provide an algorithm for splitting the initial sample in a random but repeatable way. The algorithm ensures that two researchers working independently with the same data will obtain the same results, even if the two researchers receive the data with different orderings of the observations (see Daraio et al. (2018) for additional details).

7 Dimension Reduction

The discussion in Sects. 3.1 and 3.2 makes clear that under appropriate regularity conditions, the FDH, VRS-DEA, and CRS-DEA estimators converge at rate n^{κ} where $\kappa = 1/(p+q)$, $2/(p+q+1)$, and $2/(p+q)$, respectively. In all three cases, the convergence rates become slower—and hence estimation error increases for a fixed sample size n—as the dimensionality (p + q) increases. This inverse relationship between dimensionality and convergence rates of estimators is well known in nonparametric statistics and econometrics, and is often called the "curse of dimensionality" after Bellman (1957). In the case of FDH and DEA estimators, increasing dimensionality ($p + q$) also affects bias, as seen in Sect. 5. It is perhaps less well appreciated, but nonetheless true that dimensionality also affects the variance of partial frontier estimators such as the order-m estimators and the order-α estimators discussed in Sects. 3.3 and 3.4.

Holding sample size n constant, increasing the number of inputs or the number of outputs necessarily results in a greater proportion of the sample

observations lying on FDH or DEA estimates of the frontier, i.e., more observations with efficiency estimates equal to one. Applied researchers using FDH or DEA estimators have long been aware of this phenomenon. A number of ad hoc "rules of thumb" for lower bounds on the sample size n in problems with p inputs and q outputs are proposed in the literature to address the problem. For example, Bowlin (1987), Golany and Roll (1989), Vassiloglou and Giokas (1990), and Homburg (2001) propose $n \geq 2(p+q)$. Banker et al. (1989), Bowlin (1998), Friedman and Sinuany (1998), and Raab and Lichty (2002) suggest $n \geq 3(p+q)$. Boussofiane et al. (1991) offer $n \geq pq$, and Dyson et al. (2001) recommend $n \geq 2pq$. Cooper et al. (2000, 2004), and Zervopoulos et al. (2012) advise $n \geq \max(pq,\ 3(p+q))$. No theoretical justification is given for any of these rules. Wilson (2018) provides evidence that the sample sizes suggested by these rules are too small to provide meaningful estimates of technical efficiency.

Wilson (2018) provides three diagnostics that can be used to warn researchers of situations where the number of inputs and outputs is excessive for a given sample size. The first diagnostic, the effective parametric sample size, is based on evaluating the number of observations m required in a parametric estimation problem (with the usual parametric convergence rate, $m^{1/2}$) to obtain estimation error of the same order that one would obtain with n observations in a nonparametric problem with convergence rate n^{κ}. Recall that for FDH, VRS-DEA, or CRS-DEA estimators, $\kappa = 1/(p+q)$, $2/(p+q+1)$ or $2/(p+q)$. To illustrate, note that Charnes et al. (1981) use the CRS-DEA estimator to examine $n = 70$ schools that use $p = 5$ inputs to produce $q = 3$ outputs. Simple calculations indicate that Charnes et al. achieve estimation error of the same order that one would attain with only 8 observations in a parametric problem. Using the VRS-DEA or FDH estimator with the Charnes et al. (1981) would result in estimation error of the same order that one would obtain in a parametric problem with 7 or 3 observations, respectively. One would likely be suspicious of estimates from a parametric problem with 8 or fewer observations, and one should be suspicious here, too. The ad hoc rules listed above suggest from 15 to 30 observations for the Charnes et al. (1981) application, which is less than half the number of observations used in their study.

With the CRS-DEA estimator, Charnes et al. (1981) obtain 19 efficiency estimates equal to 1. The VRS-DEA and FDH estimators yield 27 and 64 estimates equal to 1, respectively.[14]

[14]Here, efficiency estimates are obtained using the full sample of 70 observations. Charnes et al. split their sample into two groups according to whether schools participated in a social experiment known as "Program Follow Through." These data are analyzed further in Sect. 8.

Recalling that the FDH, VRS-DEA, and CRS-DEA estimators are progressively more restrictive, it is apparent that much of the inefficiency reported by Charnes et al. is due to their assumption of CRS. Under the assumptions of the statistical model (e.g., under Assumptions 2.4 and 2.5), there is not probability mass along the frontier, and hence, there is zero probability of obtaining a firm with no inefficiency. The second diagnostic suggested by Wilson (2018c) involves considering the number of observations that yield FDH efficiency estimates equal to 1. If this number is more than 25–50% of the sample size, then dimensionality may be excessive.

The third diagnostic proposed by Wilson (2018c) is related to the method for reducing the dimensionality proposed by Mouchart and Simar (2002) and Daraio and Simar (2007a, pp. 148–150). As in Sect. 3.2, let $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote the $(p \times n)$ and $(q \times n)$ matrices of observed, input and output vectors in the sample $\mathcal{S}_n$ and suppose the rows of $\boldsymbol{X}$ and $\boldsymbol{Y}$ have been standardized by dividing each element in each row by the standard deviation of the values in each row. This affects neither the efficiency measure defined in (3) nor its DEA and FDH estimators, which are invariant to units of measurement. Now consider the $(p \times p)$ and $(q \times q)$ moment matrices $\boldsymbol{X}\boldsymbol{X}'$ and $\boldsymbol{Y}\boldsymbol{Y}'$. The moment matrices are by construction square, symmetric, and positive definite. Let $\lambda_{x1}, \ldots, \lambda_{xp}$ denote the eigenvalues of $\boldsymbol{X}\boldsymbol{X}'$ arranged in decreasing order, and let Λ_x denote the $(p \times p)$ matrix whose jth column contains the eigenvector corresponding to λ_{xj}. Define

$$R_x := \frac{\lambda_{x1}}{\sum_{j=1}^{p} \lambda_{xj}}, \tag{134}$$

and use the eigenvalues of $\boldsymbol{Y}\boldsymbol{Y}'$ to similarly define R_y. Let Λ_y denote the $(q \times q)$ matrix whose jth column contains the eigenvector corresponding to λ_{yj}, the jth largest eigenvalue of $\boldsymbol{Y}\boldsymbol{Y}'$.

It is well known that R_x and R_y provide measures of how close the corresponding moment matrices are to rank one, regardless of the joint distributions of inputs and outputs. For example, if $R_x = 0.9$, then the first principal component $\boldsymbol{X}\Lambda_{x1}$ (where Λ_{x1} is the first column of Λ_x) contains 90% of the independent linear information contained in the p columns of $\boldsymbol{X}$. One might reasonably replace the p inputs with this principal component, thereby reducing dimensionality from $(p + q)$ to $(1 + q)$. If, in addition, $R_y = 0.9$, then the first principal component $\boldsymbol{Y}\Lambda_{y1}$ (where Λ_{y1} is the first column of Λ_y) contains 90% of the independent linear information contained in the q columns of $\boldsymbol{Y}$. As with inputs, one might reasonably replace the q outputs with this principle component, further reducing

dimensionality of the problem to 2. Simulation results provided by Wilson (2018) suggest that in many cases, expected estimation error will be reduced when dimensionality is reduced from $(p + q)$ to 2. Färe and Lovell (1988) show that any aggregation of inputs our outputs will distort the true radial efficiency if the technology is not homothetic, but in experiments #1 and #2 of Wilson (2018), the technology is not homothetic, but any distortion is outweighed by the reduction in estimation error resulting from dimension reduction in many cases (see Wilson (2018) for specific guidelines).

Adler and Golany (2001, 2007) propose an alternative method for reducing dimensionality. In their approach, correlation matrices of *either* inputs or outputs (but not both) are decomposed using eigensystem techniques. When estimating efficiency in the output orientation, the dimensionality of the inputs can be reduced, or when estimating efficiency in the input direction, the dimensionality of the outputs can be reduced. The Adler and Golany method cannot be used when the hyperbolic efficiency measure is estimated, and can only be used with directional efficiency in special cases. The approach of Wilson (2018) can be used in all cases. Moreover, it is well known and is confirmed by the simulation results provided by Wilson (2018) that correlation is not a particularly meaningful measure of association when data are not multivariate normally distributed, as is often the case with production data. For purposes of estimating *radial* efficiency measures such as those defined in (3), (5), and (7), the issues are not whether the data are linearly related or how they are dispersed around a central point (as measured by central moments and correlation coefficients based on central moments), but rather how similar are the rays from the origin to each datum in $\mathbb{R}_{+}^{p+q}$ (as measured by the raw moments used by Wilson [2018]).

8 An Empirical Illustration

In this section, we illustrate some of the methods described in previous sections using the "Program Follow Through" data examined by Charnes et al. (1981). All of the computations described in this section are made using the *R* programming language and the FEAR software library described by Wilson (2008).

Charnes et al. (1981) give (in Tables 1, 2, 3 and 4) observations on $p = 5$ inputs and $q = 3$ outputs for $n = 70$ schools. The first $n_1 = 49$ of these schools participated in the social experiment known as Program Follow Through, while the remaining $n_2 = 21$ schools did not. In Table 5, Charnes et al. report input-oriented CRS-DEA estimates (rounded to two

Table 1 Efficiency estimates for Charnes et al. (1981) data

Obs.	CRS-DEA	VRS-DEA	FDH	Order-α	Order-*m*
1	1.0000	1.0000	1.0000	1.0000	1.0000
2	0.9017	0.9121	1.0000	1.0000	1.0000
3	0.9883	1.0000	1.0000	1.0000	1.0000
4	0.9024	0.9035	0.9511	1.0000	0.9523
5	1.0000	1.0000	1.0000	1.7488	1.0283
6	0.9069	0.9456	1.0000	1.2500	1.0125
7	0.8924	0.8929	1.0000	1.1399	1.0018
8	0.9148	0.9192	1.0000	1.0000	1.0000
9	0.8711	0.8877	1.0000	1.0000	1.0000
10	1.0000	1.0000	1.0000	1.0000	1.0000
11	0.9819	1.0000	1.0000	1.0000	1.0000
12	0.9744	1.0000	1.0000	1.0000	1.0000
13	0.8600	0.8630	0.9866	0.9866	0.9866
14	0.9840	1.0000	1.0000	1.9429	1.0449
15	1.0000	1.0000	1.0000	3.8089	1.0343
16	0.9503	0.9507	1.0000	1.0000	1.0000
17	1.0000	1.0000	1.0000	1.7107	1.0126
18	1.0000	1.0000	1.0000	1.0000	1.0000
19	0.9501	0.9577	1.0000	1.0000	1.0000
20	1.0000	1.0000	1.0000	1.0000	1.0000
21	1.0000	1.0000	1.0000	1.0000	1.0000
22	1.0000	1.0000	1.0000	1.2081	1.0009
23	0.9630	0.9771	1.0000	1.0000	1.0000
24	1.0000	1.0000	1.0000	1.4075	1.0000
25	0.9764	0.9864	1.0000	1.1242	1.0006
26	0.9371	0.9425	1.0000	1.0000	1.0000
27	1.0000	1.0000	1.0000	1.0000	1.0000
28	0.9443	0.9903	1.0000	1.3333	1.0025
29	0.8417	0.9325	1.0000	1.3748	1.0155
30	0.9025	0.9119	1.0000	1.3093	1.0059
31	0.8392	0.8520	0.9915	1.0000	0.9918
32	0.9070	1.0000	1.0000	1.7778	1.0363
33	0.9402	0.9578	1.0000	1.0000	1.0000
34	0.8521	0.8645	1.0000	1.0000	1.0000
35	1.0000	1.0000	1.0000	1.0000	1.0000
36	0.8032	0.8033	1.0000	1.0064	1.0007
37	0.8614	0.8692	1.0000	1.0000	1.0000
38	0.9485	1.0000	1.0000	2.0205	1.0363
39	0.9352	0.9438	1.0000	1.0000	1.0000
40	1.0000	1.0000	1.0000	1.4633	1.0016
41	0.9468	0.9526	1.0000	1.1545	1.0000
42	0.9474	0.9531	1.0000	1.3333	1.0025
43	0.8708	0.8752	1.0000	1.0000	1.0000
44	1.0000	1.0000	1.0000	1.0000	1.0000
45	0.8916	1.0000	1.0000	2.0000	1.0471
46	0.9087	0.9283	1.0000	1.0000	1.0000
47	1.0000	1.0000	1.0000	1.0000	1.0000
48	1.0000	1.0000	1.0000	2.2037	1.0305

(continued)

Table 1 (continued)

Obs.	CRS-DEA	VRS-DEA	FDH	Order-α	Order-*m*
49	1.0000	1.0000	1.0000	2.2885	1.0172
50	0.9575	0.9587	1.0000	1.0000	1.0000
51	0.9205	0.9277	1.0000	1.0000	1.0000
52	1.0000	1.0000	1.0000	1.0000	1.0000
53	0.8768	0.8970	0.9872	0.9872	0.9872
54	1.0000	1.0000	1.0000	1.0000	1.0000
55	1.0000	1.0000	1.0000	1.0000	1.0000
56	1.0000	1.0000	1.0000	1.0000	1.0000
57	0.9260	0.9269	1.0000	1.0000	1.0000
58	1.0000	1.0000	1.0000	1.0000	1.0000
59	0.9223	1.0000	1.0000	1.0000	1.0000
60	0.9815	0.9981	1.0000	1.0000	1.0000
61	0.8818	0.9012	1.0000	1.0000	1.0000
62	1.0000	1.0000	1.0000	1.0000	1.0030
63	0.9611	0.9634	1.0000	1.0000	1.0000
64	0.9168	0.9373	1.0000	1.0000	1.0000
65	0.9775	0.9775	1.0000	1.0000	1.0000
66	0.9259	0.9412	0.9761	0.9761	0.9761
67	0.9271	0.9492	1.0000	1.0000	1.0000
68	1.0000	1.0000	1.0000	1.0000	1.0000
69	1.0000	1.0000	1.0000	1.0000	1.0000
70	0.9475	0.9640	1.0000	1.0000	1.0000

Table 2 Eigenvectors of input and output moment matrices

	(1)	(2)	(3)
Λ_{x1}	0.3829	0.3855	0.3799
	0.4662	0.4538	0.4493
	0.4700	0.4479	0.4541
	0.4661	0.4635	0.4394
	0.4450	0.4796	0.5046
R_x	94.9235	96.6599	93.4648
Λ_{y1}	0.5602	0.5702	0.5647
	0.5593	0.5771	0.5548
	0.6110	0.5846	0.6110
R_y	99.0422	99.1462	99.3443

decimal places) for the two groups of schools obtained by separating the schools into two groups depending on their participation in Program Follow Through and estimating efficiency for each group independent of the other group.

The implicit assumption of CRS by Charnes et al. (1981) is a strong assumption. Table 1 shows CRS-DEA, VRS-DEA, and FDH estimates of input-oriented efficiency for the Charnes et al. (1981) data. The observation

Table 3 Efficiency estimates from transformed data, split sample

Obs.	CRS-DEA	VRS-DEA	FDH	Order-α	Order-*m*
1	0.7830	0.8332	1.0000	1.0000	1.0000
2	0.6975	0.7037	0.7804	0.7821	0.7804
3	0.8212	0.8351	0.9466	0.9466	0.9466
4	0.5882	0.6285	0.6642	0.7934	0.6705
5	0.6991	0.9865	1.0000	1.2670	1.0106
6	0.7664	0.8357	0.9485	1.0000	0.9502
7	0.5560	0.5892	0.7035	0.8026	0.7075
8	0.5876	0.5927	0.6564	0.6578	0.6564
9	0.6478	0.6550	0.7449	0.7466	0.7450
10	0.8031	0.8425	1.0000	1.0000	1.0000
11	0.7765	0.7923	0.8753	0.8753	0.8753
12	0.9082	0.9300	0.9988	0.9988	0.9988
13	0.6426	0.6483	0.7203	0.7219	0.7203
14	0.7434	0.8748	0.8869	1.0000	0.8939
15	0.8987	0.9652	1.0000	1.0540	1.0013
16	0.6882	0.7022	0.7753	0.7753	0.7753
17	0.9038	0.9631	1.0000	1.1945	1.0066
18	0.9048	0.9119	1.0000	1.0000	1.0000
19	0.7281	0.7556	0.9912	0.9912	0.9912
20	1.0000	1.0000	1.0000	1.0000	1.0000
21	0.9499	0.9788	1.0000	1.0000	1.0000
22	0.9248	0.9622	1.0000	1.1408	1.0037
23	0.7470	0.7792	0.9752	0.9752	0.9752
24	0.9657	0.9872	1.0000	1.2599	1.0041
25	0.7320	0.7483	0.9544	0.9565	0.9544
26	0.7275	0.7379	0.8523	0.8523	0.8523
27	0.9016	0.9020	0.9070	0.9070	0.9070
28	0.7217	0.7513	0.7846	0.8950	0.7860
29	0.5767	0.8218	0.8395	1.0000	0.8469
30	0.6342	0.6902	0.7749	0.8167	0.7775
31	0.5660	0.6156	0.6884	0.7255	0.6899
32	0.6149	1.0000	1.0000	1.3282	1.0120
33	0.6992	0.7314	0.8925	0.8925	0.8925
34	0.6849	0.6931	0.8146	0.8146	0.8146
35	0.6809	0.7052	0.9412	0.9412	0.9412
36	0.5122	0.5617	0.6583	0.6940	0.6606
37	0.7000	0.7447	0.9137	1.0000	0.9153
38	0.8601	1.0000	1.0000	1.5371	1.0115
39	0.7344	0.7350	0.7422	0.7422	0.7422
40	0.7668	0.7927	0.8894	1.0000	0.8913
41	0.6860	0.6975	0.8429	0.8449	0.8430
42	0.7554	0.7829	0.8983	1.0000	0.8992
43	0.6502	0.6582	0.7593	0.7610	0.7593
44	0.9147	1.0000	1.0000	1.0000	1.0000
45	0.7882	0.9108	1.0000	1.4389	1.0223
46	0.6171	0.6299	0.6937	0.6937	0.6937
47	0.7679	0.7691	0.7839	0.7839	0.7843
48	0.7646	0.9111	0.9365	1.0544	0.9389

(continued)

Table 3 (continued)

Obs.	CRS-DEA	VRS-DEA	FDH	Order-α	Order-*m*
49	0.7276	0.7880	0.8605	0.9070	0.8629
50	0.6235	0.7428	0.8968	0.8968	0.8968
51	0.6003	0.6483	0.9650	0.9650	0.9650
52	0.8149	1.0000	1.0000	1.0000	1.0000
53	0.6071	0.6104	0.6413	0.6413	0.6413
54	0.7412	0.9461	1.0000	1.0000	1.0000
55	0.8248	0.8744	1.0000	1.0000	1.0000
56	0.6846	0.7058	0.9021	0.9021	0.9021
57	0.6398	0.6962	1.0000	1.0000	1.0000
58	1.0000	1.0000	1.0000	1.0000	1.0000
59	0.7205	1.0000	1.0000	1.0000	1.0000
60	0.6700	0.6901	0.8763	0.8763	0.8763
61	0.5146	0.6481	0.6481	0.8531	0.6489
62	0.7948	0.8934	1.0000	1.0000	1.0019
63	0.6240	0.6645	1.0000	1.0000	1.0000
64	0.5508	0.5703	0.7119	0.7119	0.7119
65	0.6509	0.6979	1.0000	1.0000	1.0000
66	0.4744	0.4791	0.5232	0.5232	0.5232
67	0.5809	0.6384	1.0000	1.0000	1.0000
68	0.6137	0.6706	1.0000	1.0000	1.0000
69	0.8511	1.0000	1.0000	1.0000	1.0000
70	0.6157	0.6453	0.9193	0.9193	0.9193

numbers in the first column correspond to those listed by Charnes et al. in their tables; schools represented by Observation Nos. 1–49 participated in Program Follow Through, while the schools corresponding to Observation Nos. 50–70 did not. The estimates shown in Table 1 reveal that the results are sensitive to what is assumed. In particular, the CRS-DEA estimates are quite different from the VRS-DEA estimates, which in turn are quite different from the FDH estimates. From the earlier discussion in Sects. 3.1 and 3.2, it is clear that both the VRS-DEA and FDH estimators remain consistent under CRS. It is equally clear that among the three estimators, the CRS-DEA estimator is the most restrictive. The results shown in Table 1 cast some doubt on the assumption of CRS and perhaps also the assumption of a convex production set.

In addition to the sensitivity of the results in Table 1 with respect to whether the CRS-DEA, VRS-DEA, or FDH estimator is used, the results reveal a more immediate problem. Among the CRS-DEA estimates shown in Table 1, 25—more than one-third of the sample—are equal to one, while 33 of the VRS-DEA estimates equal to one and 65 of the FDH estimates are equal to one. As discussed in Sect. 7 and by Wilson (2018), this

Table 4 Efficiency estimates from transformed data, full sample

Obs.	CRS-DEA	VRS-DEA	FDH	Order-α	Order-*m*
1	0.6733	0.8277	0.9555	0.9555	0.9555
2	0.5997	0.6640	0.7804	0.7821	0.7808
3	0.7061	0.8202	0.9466	0.9466	0.9466
4	0.5057	0.5181	0.6328	0.7934	0.6368
5	0.6011	0.7695	0.7695	1.1865	0.8049
6	0.6590	0.6876	0.9485	0.9997	0.9496
7	0.4780	0.4861	0.5611	0.7893	0.5697
8	0.5053	0.5596	0.6564	0.6578	0.6564
9	0.5570	0.6124	0.7449	0.7466	0.7449
10	0.6905	0.8341	1.0000	1.0000	1.0000
11	0.6677	0.7788	0.8753	0.8753	0.8753
12	0.7809	0.9151	0.9988	0.9988	0.9988
13	0.5525	0.6114	0.7203	0.7219	0.7203
14	0.6392	0.7102	0.8869	1.2083	0.9043
15	0.7728	0.7953	1.0000	1.0540	1.0016
16	0.5917	0.6903	0.7753	0.7753	0.7753
17	0.7771	0.7942	0.9527	1.1945	0.9561
18	0.7780	0.8639	1.0000	1.0023	1.0000
19	0.6261	0.7459	0.8079	0.8079	0.8079
20	0.8598	0.9778	1.0000	1.0000	1.0000
21	0.8168	0.9646	1.0000	1.0000	1.0000
22	0.7951	0.7954	0.7975	1.1219	0.8057
23	0.6423	0.7704	0.7949	0.7949	0.7949
24	0.8303	0.8808	1.0000	1.2599	1.0106
25	0.6294	0.6679	0.9544	0.9565	0.9547
26	0.6255	0.7243	0.8523	0.8523	0.8523
27	0.7753	0.8805	0.9070	1.0000	0.9079
28	0.6205	0.6210	0.6257	0.8802	0.6311
29	0.4958	0.6460	0.6460	0.9961	0.6734
30	0.5453	0.5680	0.7749	0.8167	0.7755
31	0.4867	0.5066	0.6884	0.7255	0.6894
32	0.5287	0.8614	0.8614	1.1338	0.8823
33	0.6012	0.7235	0.8925	0.8925	0.8925
34	0.5889	0.6799	0.8146	0.8146	0.8146
35	0.5855	0.6958	0.7672	0.7672	0.7672
36	0.4404	0.4619	0.6583	0.6938	0.6589
37	0.6019	0.6142	0.7287	1.0000	0.7362
38	0.7395	0.8178	1.0000	1.4118	1.0334
39	0.6314	0.7164	0.7422	0.8183	0.7422
40	0.6593	0.6738	0.8746	0.8894	0.8761
41	0.5898	0.6368	0.8429	0.8449	0.8430
42	0.6495	0.6578	0.8834	0.8983	0.8851
43	0.5591	0.6121	0.7593	0.7610	0.7593
44	0.7865	1.0000	1.0000	1.0000	1.0000
45	0.6778	0.7452	1.0000	1.3175	1.0344
46	0.5306	0.6193	0.6937	0.6937	0.6937
47	0.6603	0.7475	0.7839	0.8642	0.7850

(continued)

Table 4 (continued)

Obs.	CRS-DEA	VRS-DEA	FDH	Order-α	Order-*m*
48	0.6574	0.7341	0.9365	1.0559	0.9469
49	0.6256	0.6488	0.8605	0.9070	0.8615
50	0.6235	0.7291	0.8073	0.8073	0.8073
51	0.6003	0.6483	0.9650	1.0507	0.9709
52	0.8149	0.9787	1.0000	1.0000	1.0000
53	0.6071	0.6104	0.6413	0.9021	0.6485
54	0.7412	0.9169	1.0000	1.0000	1.0000
55	0.8248	0.8687	1.0000	1.0168	1.0008
56	0.6846	0.7058	0.8984	0.9469	0.8993
57	0.6398	0.6897	0.9197	0.9217	0.9205
58	1.0000	1.0000	1.0000	1.4067	1.0133
59	0.7205	1.0000	1.0000	1.0000	1.0000
60	0.6700	0.6901	0.8728	0.9199	0.8736
61	0.5146	0.6481	0.6481	1.0000	0.6789
62	0.7948	0.8934	1.0000	1.3189	1.0327
63	0.6240	0.6645	1.0000	1.0404	1.0050
64	0.5508	0.5681	0.7119	0.7239	0.7125
65	0.6509	0.6979	1.0000	1.1282	1.0113
66	0.4744	0.4791	0.5232	0.7359	0.5296
67	0.5809	0.6318	0.8088	0.8107	0.8089
68	0.6137	0.6640	0.8706	0.8726	0.8706
69	0.8511	1.0000	1.0000	1.6466	1.0481
70	0.6157	0.6453	0.9155	0.9650	0.9164

is indicative of too many dimensions for the given number of observations. Moreover, even with the assumption of CRS, the convergence rate of the CRS-DEA estimator used by Charnes et al. (1981) is $n^{1/4}$. As discussed in Sect. 7 and Wilson (2018), this results in an effective parametric sample size of 8. Moreover, an eigensystem analysis on the full data yields $R_x = 94.923$ and $R_y = 99.042$. The discussion in Sect. 7 and in Wilson (2018) makes clear the need for dimension reduction. The simulation results obtained by Wilson (2018) suggest that mean-square error is likely reduced when the 8 outputs are combined into a single principal component and the 3 outputs are combined into a single principal component using eigenvectors of the moment matrices of the inputs and outputs as described by Wilson (2018).

Table 1 also shows (input-oriented) order-α efficiency estimates (with $\alpha = 0.95$) and order-m efficiency estimates (with $m = 115$, chosen to give the number of observations *above* the order-m partial frontier similar to the number of observations lying above the order-α frontier). Note that only 3 observations lie *below* the estimated order-α frontier, and only 4 observations lie below the estimated order-m frontier (indicated by estimates less than 1). This provides further evidence that the number of dimensions is too large for the available amount of data.

Table 2 shows the eigenvectors corresponding to the largest eigenvalues for the moment matrices of inputs and outputs as discussed in Sect. 7. The first column of Table 2 gives the eigenvectors as well as values of R_x and R_y computed from the full sample with 70 observations. The second column gives similar information computed from Observation Nos. 1–49 corresponding to the Program Follow Through participants, while the third column gives similar information computed from Observation Nos. 50–70 for schools that did not participate in Program Follow Through. The results are similar across the three columns of Table 2, and so it appears reasonable to use the eigenvectors computed from the full sample to compute principal components. Doing so for both inputs and outputs reduces dimensionality from 8 to 2.

The first principal components of inputs and outputs based on moment matrices as described above are plotted in Fig. 1. The plot in Fig. 1 reveals two dissimilar observations. The observation lying in the upper right corner of the plot is Observation No. 59, and the observation lying at (105.93, 118.93) is Observation No. 44. Both of these observations are flagged by Wilson (1993, 1995) and Simar (2003) as outliers. Note that the transformation from the original $(p + q)$-dimensional space of inputs and outputs to the 2-dimensional space of first principal components amounts to an affine

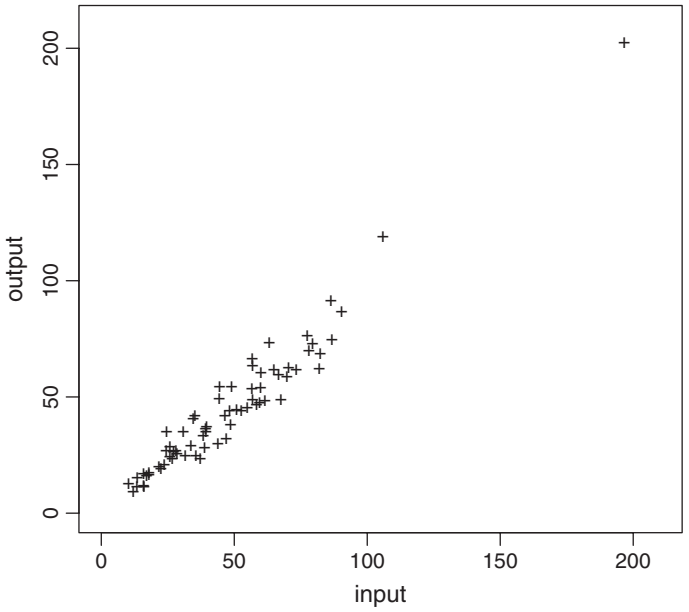

Fig. 1 Charnes et al. (1981) data after principal component transformation

function. It is well known that the image of a convex set under an affine function is also convex, and that the affine transformation preserves extreme points (e.g., see Boyd and Vandenberghe 2004). Therefore, the method for dimension reduction discussed in Sect. 7 and Wilson (2018) has an additional benefit. Reducing dimensionality to only 2 dimensions permits simple scatter plots of the transformed data, which can reveal outliers in the data that are more difficult to detect in higher dimensions.

Applying the CRS-DEA, VRS-DEA, FDH, order-α, and order-m estimators to the transformed data yields the results shown in Tables 3 and 4. The estimates in Table 3 are obtained by applying the estimators separately on the two groups of observations (i.e., those participating in Program Follow Through, corresponding to Observation Nos. 1–49, and those not participating, corresponding to Observation Nos. 50–70). The estimates in Table 4 are obtained from the full sample, ignoring program participation. In both Tables 3 and 4, the partial efficiency estimates are computed with $\alpha = 0.95$ and $m = 115$ as in Table 1.

As an illustration, consider the results for Observation No. 5 in Table 3. The CRS-DEA and VRS-DEA estimates suggest that this school could produce the same level of outputs using only 69.91% or 98.65% of its observed levels of inputs, respectively. The FDH estimate indicates that School No. 5 is on the estimated frontier, and by itself does not suggest that input levels could be reduced without reducing output levels. At the same time, however, the FDH estimates, as well as the CRS-DEA and VRS-DEA estimates, are biased upward toward 1. The order-m estimate, by contrast, suggests that School No. 5 would have to increase its input usage by 1.06% in order to meet the estimated *expected* input usage among 115 randomly chosen schools producing at least as much output as School No. 5. Similarly, the order-α estimate suggests that School No. 5 would have to increase its input usage by 26.70% while holding output levels fixed in order to have a probability of 0.05 of being dominated by a school producing more output while using less input that School No. 5. In the input orientation, the partial efficiency estimators are necessarily weakly greater than the FDH estimates. After reducing dimensionality to only two dimensions, the partial efficiency estimators have the same convergence rate (i.e., $n^{1/2}$) as the FDH estimators, but the partial efficiency estimates remain less sensitive to outliers than the FDH estimates.

Comparing results across the five columns in either Table 3 or Table 4 reveals substantial differences between estimates obtained with different estimators. This raises the question of whether CRS (assumed by Charnes et al.) is an appropriate assumption supported by the data, or whether even

convexity of the production set is an appropriate assumption. In addition, comparing estimates in Table 3 with corresponding estimates in Table 4 reveals some large differences, but not in all cases. This raises the question of whether the two groups of schools have the same mean efficiency or whether they face the same frontier. Charnes et al. allow for different frontiers, but if the schools face the same frontier, using the full sample for estimation would reduce estimation error. Rather than making arbitrary assumptions, the statistical results described in earlier sections can be used to let the data speak to what is appropriate.

In order to address the question of whether the two groups of schools face the same frontier, we apply the test of separability developed by Daraio et al. (2018) using the reduced-dimensional, transformed data. Note that here, the "environmental" variable is binary, i.e., schools either participate in Program Follow Through or they do not. Hence, no bandwidths are needed; instead, we follow the method outlined in Daraio et al. (2018, separate Appendix C) and using the difference-in-means test described by Kneip et al. (2016). To employ the test, we first randomly sort the observations using the randomization algorithm appearing in Daraio t al. (2018, separate Appendix D). We then divide the observations into two subsamples, taking the first 35 randomly sorted observations as group 1, and the remaining observations as group 2. Using the observations in group 1, we estimate efficiency using the input-oriented FDH estimator and compute the sample mean (0.8948) and sample variance (0.01424) of these estimates. We then estimate the bias (0.2540) of this sample mean using the generalized jackknife method described by Daraio et al. (2018).

For the 35 observations in group 2, we have 23 observations on schools participating in Program Follow Through (group 2a), and 12 observations not participating (group 2b). We again use the FDH estimator to estimate input-oriented efficiency, but we apply the estimator as well as the generalized jackknife independently on the two subgroups (2a and 2b). We then compute the sample mean (across 35 observations in group 2) of the efficiency estimates (0.9170), the sample variance (0.01408) and the bias correction (0.2011) as described by Daraio et al. (2018, separate Appendix C). This yields a value of 2.6278 for the test statistic given by the input-oriented analog of (124), and hence, the p-value is 0.004298.[15] Hence, the null

[15]The test is a one-sided test, since by construction the mean of the input-oriented efficiency estimates for group 1 is less than the similar mean for group 2 under departures from the null hypothesis of separability.

hypothesis of separability is soundly rejected by the data. The data provide evidence suggesting that Program Follow Through schools face a different frontier than the non-Program Follow Through schools.

We use the FDH estimator to test separability since the FDH estimator does not require an assumption of convexity, which has not yet been tested. Given that separability is rejected, we next apply the convexity test of Kneip et al. (2016) separately and independently on the 49 participating schools and the 21 non-participating schools, again using the two-dimensional, transformed data. For the participating and non-participating schools, we obtain values 3.4667 and −0.6576, respectively of the test statistic given in (112). The corresponding p-values are 0.0003 and 0.7446, and hence, we firmly reject convexity of the production set for participating schools, but we fail to reject convexity for the non-participating schools.

Next, for the non-participating schools, we test CRS against the alternative hypothesis of variable returns to scale using the method of Kneip et al. (2016). Since convexity is rejected for the Program Follow Through schools, there is no reason to test CRS. As with the other tests, we also work here again with the two-dimensional, transformed data. We obtain the value −1.7399 for the input-oriented analog of the test statistic in (119), with a corresponding p value of 0.9591. Hence, we do not reject CRS for the non-participating schools.

Taken together, the results of the tests here suggest that efficiency should be estimated separately and independently for the group of participating and the group of non-participating schools. In addition, the results indicate that the FDH estimator should be used to estimate efficiency for the participating schools. On the other hand, we do not find evidence to suggest that the CRS-DEA estimator is inappropriate for the non-participating schools.

9 Nonparametric Panel Data Frontier

9.1 Basic Ideas: Conditioning on Time

One possible approach, developed in Mastromarco and Simar (2015) in a macroeconomic setup, is to extend the basic ideas of Cazals et al. (2002) and Daraio and Simar (2005), described in Sect. 4, to a dynamic framework to allow introduction of a time dimension.

Consider a generic input vector $X \in \mathbb{R}^p_+$, a generic output vector $Y \in \mathbb{R}^q_+$ and denote by $Z \in \mathbb{R}^d$ the generic vector of environmental variables. Mastromarco and Simar (2015) consider the time T as a conditioning

variable and define for each time period t, Ψ_t, the attainable set at time t as the set of combinations of inputs and outputs feasible at time t. $\Psi_t \subset \mathbb{R}_+^{p+q}$ is the support of (X, Y) at time t, whose distribution is completely determined by

$$H_{X,Y}^t(x, y) = \text{Prob}(X \leq x, Y \geq y | T = t), \tag{135}$$

which is the probability of being dominated for a production plan (x, y) at time t. Finally, when considering the presence of additional external factors Z, the attainable set is defined as $\Psi_t^z \subseteq \Psi_t \subset \mathbb{R}_+^{p+q}$ defined as the support of the conditional probability

$$H_{X,Y|Z}^t(x, y|z) = \text{Prob}(X \leq x, Y \geq y | Z = z, T = t). \tag{136}$$

In this framework, assuming free disposability of inputs and outputs, the conditional output-oriented Debreu–Farrell technical efficiency of a production plan $(x, y) \in \Psi_t^z$, at time t facing conditions z, is defined in (54)–(56) as

$$\begin{aligned} \lambda_t(x, y|z) &= \sup\left\{ \lambda | (x, \lambda y) \in \Psi_t^z \right\} = \sup\left\{ \lambda | H_{Y,X|Z}^t(x, \lambda y|z) > 0 \right\} \\ &= \sup\left\{ \lambda | S_{Y|X,Z}^t(\lambda y|x, z) > 0 \right\}. \end{aligned} \tag{137}$$

where $S_{Y|X,Z}^t(y|x, z) = \text{Prob}(Y \geq y | X \leq x, Z = z, T = t)$.[16]

Suppose we have a sample $\mathcal{X}_{n,s} = \{(X_{it}, Y_{it}, Z_{it})\}_{i=1,\, t=1}^{n,s}$ comprised of panel data for n firms observed over s periods. Then, the unconditional and conditional attainable sets can be estimated. Assuming that the true attainable sets are convex and under free disposability of inputs and outputs (similar to Daraio and Simar 2007b), the DEA estimators at time t and facing the condition $Z = z$ is given by

$$\hat{\Psi}_{\text{DEA},t}^z = \left\{ (x, y) \in \mathbb{R}_+^p \times \mathbb{R}_+^q \,\middle|\, y \leq \sum_{j \in \mathcal{J}(z,t)} \gamma_j Y_j, x \geq \sum_{j \in \mathcal{J}(z,t)} \gamma_j X_j, \gamma \geq 0, \sum_{j \in \mathcal{J}(z,t)} \gamma_j = 1 \right\} \tag{138}$$

where $\mathcal{J}(z, t) = \left\{ j = (i, v) | z - h_z < Z_{i,v} < z + h_z; t - h_t < v < t + h_t \right\}$ and h_z and h_t are bandwidths of appropriate size selected by data-driven

[16]For efficiency measures, we only focus the presentation on the output orientation; the same could be done for any other orientation (input, hyperbolic, directional distance) (see Daraio and Simar 2007a; Bădin et al. (2010, 2012); Wilson 2011; Simar and Vanhems 2012; and Simar et al. 2012).

methods. The set $\mathcal{J}(z,t)$ describes the localizing procedure to estimate the conditional DEA estimates and determines the observations in a neighborhood of (z,t) that will be used to compute the local DEA estimate. Here, only the variables (t,z) require smoothing and appropriate bandwidths (see Sect. 4.2 for details and Daraio et al. (2018) for discussion of practical aspects of bandwidth selection).

The estimator of the output conditional efficiency measure at time t is then obtained by substituting $\hat{\Psi}^z_{DEA,t}$ for Ψ^z_t in (137). In practice, an estimate is computed by solving the linear program

$$\hat{\lambda}_{\mathrm{DEA},t}(x,y|z) = \sup\left\{\lambda|\lambda y \leq \sum_{j\in\mathcal{J}(z,t)} \gamma_j Y_j,\ x \geq \sum_{j\in\mathcal{J}(z,t)} \gamma_j x_j,\ \gamma \geq 0,\ \sum_{j\in\mathcal{J}(z,t)} \gamma_j = 1\right\} \quad (139)$$

If the convexity of the sets Ψ^z_t is questionable, it is better to use the FDH approach described in Sect. 4.2 relying only on the free disposability of the inputs and outputs. It particularizes here as follows:

$$\hat{\Psi}^z_{\mathrm{FDH},t} = \left\{(x,y) \in \mathbb{R}^p_+ \times \mathbb{R}^q_+ \middle| y \leq Y_j,\ x \geq X_j,\ j \in \mathcal{I}(z,t)\right\} \quad (140)$$

where $\mathcal{I}(z,t) = \left\{j = (i,v)|z - h_z < Z_{i,v} < z + h_z; t - h_t < v < t + h_t \cap X_{i,v} \leq x\right\}$. The conditional output FDH estimator turns out to be simply defined as (see (62)),

$$\hat{\lambda}_{\mathrm{FDH},t}(x,y|z) = \max_{i\in\mathcal{I}(z,t)} \left(\min_{j=1,\ldots,p} \left(\frac{Y^j_i}{y^j}\right)\right), \quad (141)$$

The statistical properties of these nonparametric estimators are well known as discussed in Sect. 4.2. To summarize, these estimators are consistent and converge to some non-degenerate limiting distributions to be approximated by bootstrap techniques. As discussed earlier, the rates of convergence become slower with increasing numbers of inputs and outputs, illustrating the curse of dimensionally. The situation is even jeopardized if the dimension of Z increases.

Having these estimators and to bring some insights on the effect of Z and T on the efficiency scores, Mastromarco and Simar (2015) analyze, in a second stage, the average behavior of $\lambda_t(X,Y|Z)$ at period t and conditionally on $Z = z$. The regression aims to analyze the behavior of $E\left(\hat{\lambda}_t(X,Y|Z=z)\right)$, where $\hat{\lambda}_t$ is either the DEA or the FDH estimators

defined above, as a function of z and t. For this purpose, they suggest to estimate the flexible nonparametric location-scale regression model

$$\hat{\lambda}_t(X_{it}, Y_{it}|Z_{it}) = \mu(Z_{it}, t) + \sigma(Z_{it}, t)\varepsilon_{it} \tag{142}$$

where $E(\varepsilon_{it}|Z_{it}, t) = 0$ and $\mathrm{VAR}(\varepsilon_{it}|Z_{it}, t) = 1$. This model captures both the location $\mu(z,t) = E\left(\hat{\lambda}_t(X_{it}, Y_{it}|Z_{it} = z)\right)$ and the scale effect $\sigma^2(z,t) = \mathrm{VAR}\left(\hat{\lambda}_t(X_{it}, Y_{it}|Z_{it} = z)\right)$. The nonparametric function $\mu(z,t)$ are usually estimated by local linear techniques, and $\sigma^2(z,t)$ by local constant techniques on the squares of the residuals obtained when estimating $\mu(z,t)$ (see Mastromarco and Simar [2015] for details).

As a by-product of the analysis, the scaled residual error term for a given particular production plan (X_{it}, Y_{it}, Z_{it}) at time t is computed as

$$\hat{\varepsilon}_{it} = \frac{\lambda_t(X_{it}, Y_{it}|Z_{it}) - \mu(Z_{it}, t)}{\sigma(Z_{it}, t)}. \tag{143}$$

This unexplained part of the conditional efficiency measure (called "pure efficiency" in Bădin et al. 2012) cleanses efficiency scores from external effects (here, time T and Z). The pure efficiency measure provides then a better indicator by which to assess the economic performance of production units over time and allows the ranking of production units facing different environmental factors at different time periods.

Mastromarco and Simar (2015) apply these panel techniques in a macroeconomic setup, where they use a dataset of 44 countries (26 OECD countries and 18 developing countries) over 1970–2007. In their macroeconomic cross-country framework, where countries are producers of output (i.e., GDP) given inputs (capital, labor; see Mastromarco and Simar [2015] for a detailed description of the data) and technology, inefficiency can be identified as the distance of the individual countries from the frontier estimated by the maximum output of the reference country regarded as the empirical counterpart of an optimal boundary of the production set. Inefficiencies generally reflect a sluggish adoption of new technologies, and thus efficiency improvement will represent productivity catch-up via technology diffusion.

Mastromarco Simar (2015) explore the channels under which FDI fosters productivity by disentangling the impact of this factor on the production process and its components: impact on the attainable production set (input–output space) and the impact on the distribution of efficiencies. In particular, they want to clarify the effect of time and foreign direct investment (FDI) on the catching-up process which is related to productivity gains and

so to productive efficiency of the countries. They review the literature in the field (both parametric and nonparametric).

The second-stage regression described above cleanses efficiency scores from external effects (time and FDI), and this enables Mastromarco and Simar (2015) to eliminate the common time factor effect, as economic cycles, in a very flexible and robust way (the location-scale model). Their pure efficiency measure provides a better indicator to assess the economic performance of production units over time, and in their macroeconomic framework, the "pure efficiency" represents a new measure of the Solow residual. They conclude that both time and FDI plays a role in the catching-up process, clarifying from an empirical point of view some theoretical debate on these issues.

Note that the approach developed in this section could also have been applied with robust versions of the DEA or FDH estimators by using rather the partial frontiers, the order-m and the order-α frontiers derived in Sect. 2.4. This is particularly useful if the cloud of data points contains extreme data points or outliers which may hide the true relationship between the variables (see, e.g., the discussion in Section 5.4 in Daraio and Simar [2007a]).

9.2 Cross-Sectional Dependence in Panel Data Frontier Analysis

When analyzing a panel of data, we might also expect some cross-sectional dependence between the units, this is particularly true for macroeconomic data but also for a panel of data on firms in a country, etc. In a macroeconomic setup, Mastromarco and Simar (2018) propose a flexible, nonparametric, two-step approach to take into account cross-sectional dependence due to common factors attributable to global shocks to the economy. It is easy to extend the approach for microeconomic data. They use again conditional measures where now they condition not only to external environmental factors (Z) but also to some estimates of a "common time factor" that is supposed to capture the correlation among the units. By conditioning on the latter, they eliminate the effect of these factors on the production process and so mitigate the problem of cross-sectional dependence. To define this common time factor, they follow the approach Pesaran (2006) and Bai (2009), where it is shown that an unobserved common time factor, ξ_t, can be consistently proxied by cross-sectional averages of inputs and the outputs at least asymptotically, as $n, s \to \infty$, and $s/n \to K$ where K is a finite

positive (≥ 0) constant. So the idea is to consider $F_t = (t, X_{\cdot t}, Y_{\cdot t})$ as a proxy for the unobserved nonlinear and complex common trending patterns.[17]

So, at this stage we have now a sample of observations $\{(X_{it}, Y_{it}, Z_{it}, F_t)\}_{i=1,t=1}^{n,s}$. For estimating the conditional measures, Mastromarco and Simar (2018) follow the approach suggested by Florens et al. (2014) which so far has been developed for univariate Y, but the multivariate case should not create any theoretical issues. This approach avoids direct estimation of the conditional survival function $S_{Y|X,Z,F_t}(y|x,z,f_t)$. As observed by Florens et al., the procedure reduces the impact of the curse of dimensionality (through the conditioning variables Z, F_t) and requires smoothing in these variables in the *center* of the data cloud instead of smoothing at the frontier where data are typically sparse and estimators are more sensitive to outliers. Moreover, the inclusion of time factor $F_t = (t, X_{\cdot t}, Y_{\cdot t})$ enables elimination of the common time factor effect in a flexible, nonparametric location-scale model. The statistical properties of the resulting frontier estimators are established by Florens et al. (2014).

Assume the data are generated by the nonparametric location-scale regression model

$$\begin{cases} X_{it} = \mu_x(Z_{it}, F_t) + \sigma_x(Z_{it}, F_t)\varepsilon_{x,it} \\ Y_{it} = \mu_y(Z_{it}, F_t) + \sigma_y(Z_{it}, F_t)\varepsilon_{y,it} \end{cases}, \tag{144}$$

where μ_x, σ_x and ε_x each have p components and, for ease of notation, the product of vectors is component-wise. So the first equation in (144) represents p relations, one for each component of X. We assume that each element of ε_x and ε_y has mean zero and standard deviation equal to 1. The model also assumes that $(\varepsilon_x, \varepsilon_y)$ is independent of (Z, F_t). This model allows the capture for any (z, f_t) and for each input $j = 1, \ldots, p$ and for the output, the locations $\mu_x^{(j)}(z,f_t) = E\left(X^{(j)}|Z = z, F_t = f_t\right)$, $\mu_y(z,f_t) = E(Y|Z = z, F_t = f_t)$ and the scale effects $\sigma_x^{(j),2}(z,f_t) = \mathrm{VAR}\left(X^{(j)}|Z = z, F_t = f_t\right)$, $\sigma_y^2(z,t) = \mathrm{VAR}(Y|Z = z, F_t = f_t)$ of the environmental and common factors on the production plans. Here again, for a vector a, $a^{(j)}$ denotes its jth component.

The production frontier can be estimated in two stages as proposed by Florens et al. (2014). In the first stage, the location functions $\mu_\ell(z_{it}, f_t)$ in (144) are estimated by local linear methods. Then, the variance functions

[17]Here, we use the standard notation where a dot in a subscript signifies an average over the corresponding index.

$\sigma_\ell^2(z_{it}, f_t)$ are estimated by regressing the squares of the residuals obtained from the first local linear regression on (z, f_t). For the variance functions, a local constant estimator is used to avoid negative values of the estimated variances.

The first-stage estimation yields the residuals

$$\hat{\varepsilon}_{x,it} = \frac{X_{it} - \hat{\mu}_x(Z_{it}, F_t)}{\hat{\sigma}_x(Z_{it}, F_t)} \tag{145}$$

and

$$\hat{\varepsilon}_{y,it} = \frac{Y_{it} - \hat{\lambda}_y(Z_{it}, F_t)}{\hat{\sigma}_y(Z_{it}, F_t)}, \tag{146}$$

where for ease of notation, a ratio of two vectors is understood to be component-wise. These residuals amount to whitened inputs and output obtained by eliminating the influence of the external and other environmental variables as common factors. To validate the location-scale model, Florens et al. (2014) propose a bootstrap-based testing procedure to test the independence between $\left(\hat{\varepsilon}_{x,it}, \hat{\varepsilon}_{y,it}\right)$ and (Z_{it}, F_t), i.e., the independence of whitened inputs and output from the external and global effects.

Note that here, for finite sample the independence between $(\varepsilon_x, \varepsilon_y)$ and of (Z, F_t) is not verified in finite samples, as, e.g., $\mathrm{COV}(Y_{\cdot t}, \varepsilon_{y,it}) = \sigma_y \mathrm{COV}(\varepsilon_{\cdot t}, \varepsilon_{y,it}) = \sigma_y \frac{\mathrm{VAR}(\varepsilon_{y,it})}{n}$, but since the latter converge to zero as $n \to \infty$, this does not contradict the asymptotic independence assumed by the model.

In the second stage, a production frontier is estimated for the whitened output and inputs given by (145) and (146). Hence, for each observation (i, t) a measure of "pure" efficiency is obtained. This approach accommodates both time and cross-sectional dependence and yields more reliable estimates of efficiency.[18] Moreover, as observed by Florens et al. (2014), by cleaning the dependence of external factors in the first stage, the curse of dimensionality due to the dimension of the external variables is avoided when estimating the production frontier. Smoothing over time accounts for the panel structure of the data as well as the correlation over time of observed units.

The attainable set of pure inputs and output $(\varepsilon_x, \varepsilon_y)$ is defined by

$$\Psi_\varepsilon = \left\{ (e_x, e_y) \in \mathbb{R}^{p+1} \middle| H_{\varepsilon_x, \varepsilon_y}\left(e_x, e_y\right) \geq 0 \right\} \tag{147}$$

[18]To some extent, the first step permits controlling for endogeneity due to reverse causation between production process of inputs and output and the external variables (Z, F_t).

where $H_{\varepsilon_x,\varepsilon_y}(e_x, e_y) = \Pr(\varepsilon_x \leq e_x, \varepsilon_y \geq e_y)$. The nonparametric FDH estimator is obtained by using the empirical analog $\hat{H}_{\varepsilon_x,\varepsilon_y}(e_x, e_y)$ of $H_{\varepsilon_x,\varepsilon_y}(e_x, e_y)$ and the observed residuals defined in (145) and (146). As shown in Florens et al. (2014), replacing the unobserved $(\varepsilon_x, \varepsilon_y)$ by their empirical counterparts $(\hat{\varepsilon}_x, \hat{\varepsilon}_y)$ does not change the usual statistical properties of frontier estimators. Consequently, both consistency for the full-frontier FDH estimator and $\sqrt{n}$-consistency and asymptotic normality for the robust order-m frontiers follow. Florens et al. (2014) conjecture that if the functions μ_ℓ and σ_ℓ for $\ell = x, y$ are smooth enough, the conditional FDH estimator keeps its usual nonparametric rate of convergence, i.e., here, $n^{1/(p+1)}$.

A "pure" measure of efficiency can be obtained by measuring the distance of a particular point $(\varepsilon_{x,it}, \varepsilon_{y,it})$ to the efficient frontier. Since the pure inputs and output are centered on zero, they may have negative values, requiring use of directional distance defined for a particular unit (e_x, e_y) by

$$\delta(e_x, e_y; d_x, d_y) = \sup\{\gamma | H_{\varepsilon_x,\varepsilon_y}(e_x - \gamma d_x, e_y + \gamma d_y) > 0\}, \quad (148)$$

where $d_x \in \mathbb{R}_+^p$ and $d_y \in \mathbb{R}_+$ are the chosen directions. In Mastromarco and Simar (2018), an output orientation is chosen by specifying $d_x = 0$ and $d_y = 1$. When only some elements of d_x are zero (see Daraio and Simar (2014) for details on practical computation).

In the output orientation and in the case of univariate output, the optimal production frontier can be described at any value of the pure input $e_x \in \mathbb{R}^p$ by the function

$$\varphi(e_x) = \sup\{e_y | H_{\varepsilon_x,\varepsilon_y}(e_x, e_y) > 0\}, \quad (149)$$

so that the distance to the frontier of a point (e_x, e_y) in the output direction is given directly by $\delta(e_x, e_y; 0, 1) = \varphi(e_x) - e_y$. Then, for each unit in the sample $\mathcal{X}_{n,s}$, the "pure" efficiency estimator is obtained through

$$\hat{\delta}(\hat{\varepsilon}_{x,it}, \hat{\varepsilon}_{y,it}; 0, 1) = \hat{\varphi}(\hat{\varepsilon}_{x,it}) - \hat{\varepsilon}_{y,it}, \quad (150)$$

where $\hat{\varphi}(\cdot)$ is the FDH estimator of the pure efficient frontier in the output direction. The latter is obtained as

$$\begin{aligned} \hat{\varphi}(e_x) &= \sup\left\{ e_y \middle| \hat{H}_{\varepsilon_x,\varepsilon_y}(e_x, e_y) > 0 \right\} \\ &= \max_{\{(i,t)|\hat{\varepsilon}_{x,it} \leq e_x\}} \hat{\varepsilon}_{y,it}. \end{aligned} \quad (151)$$

Similar expressions can be derived for the order-m efficiency estimator. The order-m frontier at an input value e_x is the expected value of the maximum of the outputs of m units drawn at random in the population of units such that $\varepsilon_{x,it} \leq e_x$. The nonparametric estimator is obtained by the empirical analog

$$\hat{\varphi}_m(e_x) = \hat{E}\left[\max\left(\varepsilon_{y,1t}, \ldots, \varepsilon_{y,mt}\right)\right], \tag{152}$$

where the $\varepsilon_{y,it}$ are drawn from the empirical conditional survival function $\hat{S}_{\varepsilon_y|\varepsilon_x}\left(e_y|\hat{\varepsilon}_{x,it} \leq e_x\right)$. This can be computed by Monte Carlo approximation or by solving a univariate integral using numerical methods (for practical details, see Simar and Vanhems 2012).

Note that it is possible to recover the conditional frontier in the original units, both for the full frontier and for the order-m one. As shown in Florens et al. (2014), it is directly obtained at any values of (x, z, f_t) as

$$\begin{aligned}\tau(x, z, f_t) &= \sup\left\{y|S_{Y|X,Z,F_t}(y|x, z, f_t) > 0\right\} \\ &= \mu_y(z, f_t) + \varphi(e_x)\sigma_y(z, f_t).\end{aligned} \tag{153}$$

which can be estimated by

$$\hat{\tau}(X_{it}, Z_{it}, f_t) = \hat{\mu}_y(Z_{it}, f_t) + \hat{\varphi}\left(\hat{\varepsilon}_{x,it}\right)\hat{\sigma}_y(Z_{it}, f_t). \tag{154}$$

As shown in Mastromarco and Simar (2018), this is equivalent to

$$\hat{\tau}(X_{it}, Z_{it}, f_t) = Y_{it} + \hat{\delta}\left(\hat{\varepsilon}_{x,it}, \hat{\varepsilon}_{y,it}; 0, 1\right)\hat{\sigma}_y(Z_{it}, f_t), \tag{155}$$

which has the nice interpretation that gap, in original units, between an observation (X_{it}, Y_{it}) facing the conditions (Z_{it}, f_t) and the efficient frontier is given by the corresponding "pure" efficiency measure rescaled by the local standard deviation $\hat{\sigma}_y(Z_{it}, f_t)$. In the same spirit, the conditional output-oriented Farrell efficiency estimate in original units is given by

$$\hat{\lambda}(X_{it}, Y_{it}|Z_{it}, f_t) = \frac{\hat{\tau}(X_{it}, Z_{it}, f_t)}{Y_{it}}. \tag{156}$$

Note that there were no need to estimate the conditional survival function $S_{Y|X,Z,F_t}(y|x, z, f_t)$ to obtain this result. Similar results are described in Mastromarco and Simar (2018) or order-m robust frontiers.

Mastromarco and Simar (2018) have applied the approach in the analysis of the productivity performance of 26 OECD countries and 18 developing countries considering the spillover effects of global shocks and business cycles due to

increasing globalization and interconnection among countries. So far all studies analyzing effect of common external factors on productivity of countries have been on the stream of parametric modeling which suffers of misspecification problem due to unknown data generation process in applied studies. The frontier model used by Mastromarco and Simar (2018) enables investigation of whether the effect of environmental/global variables on productivity occurs via technology change or efficiency. Mastromarco and Simar (2018) quantify the impact of environmental/global factors on efficiency levels and make inferences about the contributions of these variables in affecting efficiency. Specifically, Mastromarco and Simar (2018) assess the impact of FDI on the production process for small, medium, and large countries. They intend to redress an important policy issue of whether the protection-oriented policy will hamper the production efficiency through limiting FDI by explicitly analyzing the relationship between efficiency and openness factor FDI dependent on size of country.

Mastromarco and Simar (2018) show that, especially for medium and big countries, FDI appears to play an important role in accelerating the technological change (shifts in the frontier) but with a decreasing effect at large values of FDI. This result confirms the theoretical hypothesis that FDI leads to increase in productivity by spurring competition (Glass and Saggi 1998). Moreover, their findings reveal that knowledge embodied in FDI is transferred for technology externalities (shift of the frontier) (Cohen and Levinthal 1989), supporting the evidence highlighting that lowering trade barriers have exerted a significantly positive effects on productivity (e.g., Borensztein et al. 1998 and Cameron et al. 2005).

In a further development in the panel data context, Mastromarco and Simar (2017) address the problem of endogeneity due to latent heterogeneity. They analyze the influence of human capital (measured as average years of education in the population) on the production process of a country by extending the instrumental nonparametric approach proposed by Simar et al. (2016). The latent factor which is identified is the part of human capital independent of the life expectancy in the countries. It appears that this latent factor can be empirically interpreted as innovation, quality of the institutions and the difference in property rights systems among countries (see Mastromarco and Simar 2017 for details).

10 Summary and Conclusions

Benchmarking production performance is a fascinating, but difficult field because of the nonstandard and challenging statistical and econometric problems that are involved. Nonparametric methods for efficiency estimation

bring together a wide variety of mathematical tools from mathematical statistics, econometrics, computer science, and operations research. As this survey indicates, a lot of progress has been made in recent years in establishing statistical theory for efficiency estimators. The development of new CLTs for both unconditional and conditional efficiency estimators opens the door to testing hypotheses about the structure of production models and their evolution over time. The extension of existing results to panel data similarly opens new doors to handle dynamic situations. These and other issues continue to be addressed by the authors of this survey as well as others.

Although we have focused on the nonparametric, deterministic framework, other possibilities fall somewhere between this framework and the framework of fully parametric, stochastic frontiers, for example, Fan et al. (1996) propose semi-parametric estimation of a stochastic frontier. Kumbhakar et al. (2007) propose a local likelihood approach that requires distributional assumptions for the noise and efficiency processes, but does not require functional-form assumptions for the frontier. Parameters of the noise and efficiency distributions are estimated locally, and hence, the method is almost fully nonparametric. In a similar vein, Simar et al. (2017) use moment-based methods to avoid specification of the efficiency process and also avoid specifying the distribution of (the symmetric) noise process. Their method has the additional advantage that from a computational viewpoint, it is much easier to implement than the method of Kumbhakar et al. (2007). Simar and Zelenyuk (2011) discuss stochastic version of the FDH and DEA estimators. Kneip et al. (2015a) allow for measurement error, while Florens et al. (2018) allow for symmetric noise with unknown variance in a nonparametric framework. In yet another direction, Kuosmanen and Kortelainen (2012) introduce shape constraints in a semi-parametric framework for frontier estimation, but so far all the stochastic parts of the model are fully parametric. All of these create new issues and new directions for future research.

References

Adler, N., and B. Golany. 2001. Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to western Europe. *European Journal of Operational Research* 132: 260–273.

———. 2007. PCA-DEA: Reducing the curse of dimensionality. In *Modeling data irregularities and structural complexities in data envelopment analysis*, eds. J. Zhu and W. D. Cook, pp. 139–154. New York: Springer Science + Business Media, LLC.

Apon, A. W., L. B. Ngo, M. E. Payne, and P. W. Wilson. 2015. Assessing the effect of high performance computing capabilities on academic research output. *Empirical Economics* 48, 283–312.

Aragon, Y., A. Daouia, and C. Thomas-Agnan. 2005. Nonparametric frontier estimation: A conditional quantile-based approach. *Econometric Theory* 21, 358–389.

Bădin, L., C. Daraio, and L. Simar. 2010. Optimal bandwidth selection for conditional efficiency measures: A data-driven approach. *European Journal of Operational Research* 201, 633–664.

———. 2012. How to measure the impact of environmental factors in a nonparametric production model. *European Journal of Operational Research* 223, 818–833.

———. 2014. Explaining inefficiency in nonparametric production models: The state of the art. *Annals of Operations Research* 214, 5–30.

Bahadur, R.R., and L.J. Savage. 1956. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics* 27: 1115–1122.

Bai, J. 2009. Panel data models with interactive fixed effects. *Econometrica* 77: 1229–1279.

Banker, R., A. Charnes, W.W. Cooper, J. Swarts, and D.A. Thomas. 1989. An introduction to data envelopment analysis with some of its models and uses. *Research in Governmental and Nonprofit Accounting* 5: 125–165.

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

Bellman, R.E. 1957. *Dynamic programming*. Princeton, NJ: Princeton University Press.

Bickel, P.J., and D.A. Freedman. 1981. Some asymptotic theory for the bootstrap. *Annals of Statistics* 9: 1196–1217.

Bickel, P.J., and A. Sakov. 2008. On the choice of *m* in the *m* out of *n* bootstrap and confidence bounds for extrema. *Statistica Sinica* 18: 967–985.

Bogetoft, P. 1996. DEA on relaxed convexity assumptions. *Management Science* 42: 457–465.

Bogetoft, P., J.M. Tama, and J. Tind. 2000. Convex input and output projections of nonconvex production possibility sets. *Management Science* 46: 858–869.

Borensztein, E., J.D. Gregorio, and L.-W. Lee. 1998. How does foreign direct investment affect economic growth. *Journal of International Economics* 45: 115–135.

Boussofiane, A., R.G. Dyson, and E. Thanassoulis. 1991. Applied data envelopment analysis. *European Journal of Operational Research* 52: 1–15.

Bowlin, W.F. 1987. Evaluating the efficiency of US Air Force real-property maintenance activities. *Journal of the Operational Research Society* 38: 127–135.

———. 1998. Measuring performance: An introduction to data envelopment analysis (DEA). *Journal of Applied Behavior Analysis* 15, 3–27.

Boyd, S., and L. Vandenberghe. 2004. *Convex optimization*. New York: Cambridge University Press.

Briec, W., K. Kerstens, and P. Van den Eeckaut. 2004. Non-convex technologies and cost functions: Definitions, duality, and nonparametric tests of convexity. *Journal of Economics* 81: 155–192.

Cameron, G., J. Proudman, and S. Redding. 2005. Technological convergence, R & D, trade and productivity growth. *European Economic Review* 49: 775–807.

Cazals, C., J.P. Florens, and L. Simar. 2002. Nonparametric frontier estimation: A robust approach. *Journal of Econometrics* 106: 1–25.

Chambers, R.G., Y. Chung, and R. Färe. 1998. Profit, directional distance functions, and Nerlovian efficiency. *Journal of Optimization Theory and Applications* 98: 351–364.

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

Charnes, A., W.W. Cooper, and E. Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2: 429–444.

———. 1981. Evaluating program and managerial efficiency: An application of data envelopment analysis to program follow through, *Management Science* 27, 668–697.

Cohen, W. and D. Levinthal. 1989. Innovation and learning: Two faces of R & D. *Economics Journal* 107, 139–149.

Cooper, W. W., S. Li, L. M. Seiford, and J. Zhu 2004, Sensitivity analysis in DEA. In *Handbook on data envelopment analysis*, eds. W. W. Cooper, L. M. Seiford, and J. Zhu, Chapter 3, pp. 75–98. New York: Kluwer Academic Publishers.

Cooper, W.W., L.M. Seiford, and K. Tone. 2000. *Data envelopment analysis: A comprehensive text with models, applications, references and DEA-solver software*. Boston, MA: Kluwer Academic Publishers.

Daouia, A., J.P. Florens, and L. Simar. 2010. Frontier estimation and extreme value theory. *Bernoulli* 16: 1039–1063.

Daouia, A., and L. Simar. 2007. Nonparametric efficiency analysis: A multivariate conditional quantile approach. *Journal of Econometrics* 140: 375–400.

Daouia, A., L. Simar, and P.W. Wilson. 2017. Measuring firm performance using nonparametric quantile-type distances. *Econometric Reviews* 36: 156–181.

Daraio, C., and L. Simar. 2005. Introducing environmental variables in nonparametric frontier models: A probabilistic approach. *Journal of Productivity Analysis* 24: 93–121.

———. 2007a. *Advanced robust and nonparametric methods in efficiency analysis*. New York: Springer Science + Business Media, LLC.

———. 2007b. Conditional nonparametric frontier models for convex and nonconvex technologies: A unifying approach. *Journal of Productivity Analysis* 28, 13–32.

———. 2014. Directional distances and their robust versions: Computational and testing issues. *European Journal of Operational Research* 237, 358–369.

Daraio, C., L. Simar, and P.W. Wilson. 2018. Central limit theorems for conditional efficiency measures and tests of the 'separability condition' in non-parametric, two-stage models of production. *The Econometrics Journal* 21: 170–191.

Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19: 273–292.

Deprins, D., L. Simar, and H. Tulkens. 1984. Measuring labor inefficiency in post offices. In *The performance of public enterprises: Concepts and measurements*, ed. M.M.P. Pestieau and H. Tulkens, 243–267. Amsterdam: North-Holland.

Durstenfeld, R. 1964. Algorithm 235: Random permutation. *Communications of the ACM* 7: 420.

Dyson, R.G., R. Allen, A.S. Camanho, V.V. Podinovski, C.S. Sarrico, and E.A. Shale. 2001. Pitfalls and protocols in DEA. *European Journal of Operational Research* 132: 245–259.

Fan, Y., Q. Li, and A. Weersink. 1996. Semiparametric estimation of stochastic production frontier models. *Journal of Business and Economic Statistics* 14: 460–468.

Färe, R. 1988. *Fundamentals of production theory*. Berlin: Springer-Verlag.

Färe, R., and S. Grosskopf. 2004. *Efficiency and productivity: new directions*. Boston, MA: Kluwer Academic Publishers.

Färe, R., S. Grosskopf, and C.A.K. Lovell. 1985. *The measurement of efficiency of production*. Boston: Kluwer-Nijhoff Publishing.

Färe, R., S. Grosskopf, and D. Margaritis. 2008. Productivity and efficiency: Malmquist and more. In *The measurement of productive efficiency*, eds. H. Fried, C. A. K. Lovell, and S. Schmidt, Chapter 5, 2nd edition, pp. 522–621. Oxford: Oxford University Press.

Färe, R., and C.A.K. Lovell. 1988. Aggregation and efficiency. In *Measurement in economics*, ed. W. Eichhorn, 639–647. Heidelberg: Physica-Verlag.

Farrell, M.J. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society a* 120: 253–281.

Fisher, R. A. and F. Yates. 1948. *Statistical tables for biological, agricultural and medical research*, 3rd edition. London: Oliver and Boyd.

Florens, J.P., L. Simar, and I. Van Keilegom. 2014. Frontier estimation in nonparametric location-scale models. *Journal of Econometrics* 178: 456–470.

———. 2018. Stochastic frontier estimation with symmetric error. Discussion paper #2018/xx, Institut de Statistique Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Friedman, L., and Z. Sinuany-Stern. 1998. Combining ranking scales and selecting variables in the DEA context: The case of industrial branches. *Computers & Operations Research* 25: 781–791.

Glass, A.J., and K. Saggi. 1998. International technological transfer and technology gap. *Journal of Development Economics* 55: 369–398.

Golany, B., and Y. Roll. 1989. An application procedure for DEA. *Omega* 17: 237–250.

Gray, H.L., and R.R. Schucany. 1972. *The Generalized Jackknife Statistic*. New York: Marcel Decker Inc.

Homburg, C. 2001. Using data envelopment analysis to benchmark activities. *International Journal of Production Economics* 73: 51–58.

Hsu, P.L., and H. Robbins. 1947. Complete convergence and the law of large numbers. *Proceedings of the National Academy of Sciences of the United States of America* 33: 25–31.

Jeong, S.O., B.U. Park, and L. Simar. 2010. Nonparametric conditional efficiency measures: Asymptotic properties. *Annals of Operations Research* 173: 105–122.

Jeong, S.O., and L. Simar. 2006. Linearly interpolated FDH efficiency score for nonconvex frontiers. *Journal of Multivariate Analysis* 97: 2141–2161.

Kneip, A., B. Park, and L. Simar. 1998. A note on the convergence of nonparametric DEA efficiency measures. *Econometric Theory* 14: 783–793.

Kneip, A., L. Simar, and I. Van Keilegom. 2015a. Frontier estimation in the presence of measurement error with unknown variance. *Journal of Econometrics* 184, 379–393.

Kneip, A., L. Simar, and P.W. Wilson. 2008. Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models. *Econometric Theory* 24: 1663–1697.

———. 2011. A computationally efficient, consistent bootstrap for inference with non-parametric DEA estimators. *Computational Economics* 38, 483–515.

———. 2015b. When bias kills the variance: Central limit theorems for DEA and FDH efficiency scores. *Econometric Theory* 31, 394–422.

———. 2016. Testing hypotheses in nonparametric models of production. *Journal of Business and Economic Statistics* 34, 435–456.

Koopmans, T. C. 1951. An analysis of production as an efficient combination of activities. In *Activity analysis of production and allocation*, ed. T. C. Koopmans, pp. 33–97. New York: Wiley. Cowles Commission for Research in Economics, Monograph 13.

Kumbhakar, S.C., B.U. Park, L. Simar, and E.G. Tsionas. 2007. Nonparametric stochastic frontiers: A local likelihood approach. *Journal of Econometrics* 137: 1–27.

Kuosmanen, T., and M. Kortelainen. 2012. Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis* 38: 11–28.

Mastromarco, C., and L. Simar. 2015. Effect of FDI and time on catching up: New insights from a conditional nonparametric frontier analysis. *Journal of Applied Econometrics* 30: 826–847.

———. 2017. Cross-section dependence and latent heterogeneity to evaluate the impact of human capital on country performance. Discussion paper #2017/30, Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

———. 2018. Globalization and productivity: A robust nonparametric world frontier analysis. *Economic Modelling* 69, 134–149.

Mouchart, M. and L. Simar. 2002. Efficiency analysis of air controllers: First insights. Consulting report #0202, Institut de Statistique, Université Catholique de Louvain, Belgium.

Park, B.U., S.-O. Jeong, and L. Simar. 2010. Asymptotic distribution of conical-hull estimators of directional edges. *Annals of Statistics* 38: 1320–1340.

Park, B.U., L. Simar, and C. Weiner. 2000. FDH efficiency scores from a stochastic point of view. *Econometric Theory* 16: 855–877.

Pesaran, M.H. 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* 74: 967–1012.

Politis, D.N., J.P. Romano, and M. Wolf. 2001. On the asymptotic theory of subsampling. *Statistica Sinica* 11: 1105–1124.

Raab, R., and R. Lichty. 2002. Identifying sub-areas that comprise a greater metropolitan area: The criterion of country relative efficiency. *Journal of Regional Science* 42: 579–594.

Shephard, R.W. 1970. *Theory of cost and production functions*. Princeton: Princeton University Press.

Simar, L. 2003. Detecting outliers in frontier models: A simple approach. *Journal of Productivity Analysis* 20: 391–424.

Simar, L., I. Van Keilegom, and V. Zelenyuk. 2017. Nonparametric least squares methods for stochastic frontier models. *Journal of Productivity Analysis* 47: 189–204.

Simar, L., and A. Vanhems. 2012. Probabilistic characterization of directional distances and their robust versions. *Journal of Econometrics* 166: 342–354.

Simar, L., A. Vanhems, and I. Van Keilegom. 2016. Unobserved heterogeneity and endogeneity in nonparametric frontier estimation. *Journal of Econometrics* 190: 360–373.

Simar, L., A. Vanhems, and P.W. Wilson. 2012. Statistical inference for DEA estimators of directional distances. *European Journal of Operational Research* 220: 853–864.

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

———. 1999a. Some problems with the Ferrier/Hirschberg bootstrap idea, *Journal of Productivity Analysis* 11, 67–80.

———. 1999b. Of course we can bootstrap DEA scores! But does it mean anything? Logic trumps wishful thinking. *Journal of Productivity Analysis* 11, 93–97.

———. 2000. A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics* 27, 779–802.

———. 2007. Estimation and inference in two-stage, semi-parametric models of productive efficiency. *Journal of Econometrics* 136, 31–64.

———. 2008. Statistical inference in nonparametric frontier models: Recent developments and perspectives. In *The Measurement of Productive Efficiency*, chapter 4,

eds. H. O. Fried, C. A. K. Lovell, and S. S. Schmidt, 2nd edition, pp. 421–521. Oxford: Oxford University Press.

———. 2011a. Inference by the *m* out of *n* bootstrap in nonparametric frontier models. *Journal of Productivity Analysis* 36, 33–53.

———. 2011b. Two-Stage DEA: Caveat emptor. *Journal of Productivity Analysis* 36, 205–218.

———. 2013. Estimation and inference in nonparametric frontier models: Recent developments and perspectives. *Foundations and Trends in Econometrics* 5, 183–337.

Simar, L., and V. Zelenyuk. 2011. Stochastic FDH/DEA estimators for frontier analysis. *Journal of Productivity Analysis* 36: 1–20.

Spanos, A. 1999. *Probability theory and statistical inference: Econometric modeling with observational data*. Cambridge: Cambridge University Press.

Varian, H.R. 1978. *Microeconomic analysis*. New York: W. W. Norton & Co.

Vassiloglou, M., and D. Giokas. 1990. A study of the relative efficiency of bank branches: An application of data envelopment analysis. *Journal of the Operational Research Society* 41: 591–597.

Wheelock, D.C., and P.W. Wilson. 2008. Non-parametric, unconditional quantile estimation for efficiency analysis with an application to Federal Reserve check processing operations. *Journal of Econometrics* 145: 209–225.

Wilson, P.W. 1993. Detecting outliers in deterministic nonparametric frontier models with multiple outputs. *Journal of Business and Economic Statistics* 11: 319–323.

———. 1995. Detecting influential observations in data envelopment analysis. *Journal of Productivity Analysis* 6, 27–45.

———. 2008. FEAR: A software package for frontier efficiency analysis with R. *Socio-Economic Planning Sciences* 42, 247–254.

———. 2011. Asymptotic properties of some non-parametric hyperbolic efficiency estimators. In *Exploring research frontiers in contemporary statistics and econometrics,* eds. I. Van Keilegom and P. W. Wilson, pp. 115–150. Berlin: Springer-Verlag.

———. 2018. Dimension reduction in nonparametric models of production. *European Journal of Operational Research* 267, 349–367.

Zervopoulos, P. D., F. Vargas, and G. Cheng. 2012. Reconceptualizing the DEA bootstrap for improved estimations in the presence of small samples. In *Data envelopment analysis: Theory and applications*, eds. R. Banker, A. Emrouznejad, A. L. M. Lopes, and M. R. de Almeida, Chapter 27, pp. 226–232. Brazil: Proceedings of the 10th International Conference on DEA.

Bayesian Performance Evaluation

Mike G. Tsionas

1 Introduction

Performance evaluation is a task of considerable importance in both Data Envelopment Analysis (DEA) and stochastic frontier models (SFM), see chapters "Ranking Methods Within Data Envelopment Analysis" and "Distributional Forms in Stochastic Frontier Analysis". In this chapter, we review techniques for performance evaluation with a focus on SFM. The standard argument in favor of DEA is that it does not assume a functional form or a distribution of error terms. While this is true, there are two problems. First, DEA cannot be applied to huge data sets as a linear programming problem must be solved for each decision-making unit (DMU). Second, the rate of convergence is slow and depends on the sum of the number of inputs and outputs which, often, is quite large. This is true despite the recent progress that has been made in DEA using the bootstrap (Simar and Wilson 1998, 2000, 2004). It is, of course, quite possible that with parallel computing in hundreds of processors, this will change in the not too distant future.

The distinguishing feature of the Bayesian approach in performance evaluation is that parameter uncertainty of the various models is *formally* taken into account along with *model uncertainty* as well. The usual sampling-

M. G. Tsionas (✉)
Lancaster University Management School, Lancaster, UK
e-mail: m.tsionas@lancaster.ac.uk

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_11

theory approach proceeds *conditionally* on the parameter estimates that have been obtained. Of course, the bootstrap can be used but the justification of the bootstrap itself is only asymptotic. We do not intend here to compare and contrast, in detail, the Bayesian versus the sampling-theory approach. We should mention, however, that the Bayesian approach is equipped with numerical techniques that can handle complicated models of performance, where the sampling-theory approach is quite difficult to implement.

2 Bayesian Performance Analysis in Stochastic Frontier Models

Consider the basic SFM in chapter "Distributional Forms in Stochastic Frontier Analysis":

$$y_i = x_i'\beta + v_i - u_i, i = 1, \ldots, n, \tag{1}$$

where x_i is a $k \times 1$ vector of inputs, β is a $k \times 1$ vector of coefficients, v_i is the two-sided error term, and $u_i \geq 0$ is the one-sided error term. Usually, x_i contains a column of ones, and the inputs and output are in logs so that u_i measures technical inefficiency in percentage terms. Suppose, for simplicity, that

$$v_i \sim iidN\left(0, \sigma_v^2\right), \tag{2}$$

and independently of v_i and x_i, the u_i s are iid and follow a distribution with density $p(u|\alpha)$ which depends on certain unknown parameters α. It is known that we can write the joint distribution for the ith observation as follows:

$$p(y_i|x_i, \theta) = \int_0^\infty \left(2\pi\sigma_v^2\right)^{-1/2} \exp\left\{-\tfrac{1}{\sigma_v^2}\left(y_i + u_i - x_i'\beta\right)^2\right\} p(u_i|\alpha)\mathrm{d}u_i, \tag{3}$$

where $\theta = \left(\beta', \sigma_v, \alpha'\right)' \in \Theta \subseteq \mathbb{R}^d$ is the parameter vector. For certain densities, like the half-normal or exponential we know from chapter "Distributional Forms in Stochastic Frontier Analysis" that the integral is available in closed form, so the likelihood function

$$L(\theta; Y) = \prod_{i=1}^n f(y_i; \theta), \tag{4}$$

is available in closed form and maximum likelihood (ML) can be applied easily to obtain parameter estimates $\widehat{\theta}$ for θ.

One of the first Bayesian approaches to the problem has been provided by van den Broeck et al. (1994) who have used importance sampling to perform the computations. A more recent discussion is provided in Assaf et al. (2017).

First of all, to introduce the Bayesian approach we need a prior $p(\theta)$ for the elements of θ. By Bayes' theorem, we know that, after observing the data, the posterior distribution is:

$$p(\theta|Y) \propto L(\theta; Y)p(\theta). \tag{5}$$

More precisely,

$$p(\theta|Y) = \frac{L(\theta; Y)p(\theta)}{\int_{\Theta} L(\theta; Y)p(\theta)d\theta}. \tag{6}$$

The denominator, $M(Y) \equiv \int_{\Theta} L(\theta; Y)p(\theta)d\theta$ is known as the marginal likelihood and provides the integrating constant of the posterior. Typically, however, we do not need the integrating constant to implement schemes for posterior inference.

In the Bayesian approach, it is typical to consider what is known as data augmentation in the statistical literature (Tanner and Wong 1987). Data augmentation works as follows. Suppose the posterior $p(\theta|Y)$ is difficult to analyze. For example, computing posterior moments or marginal densities boils down to computing certain intractable integrals of the form $\int_{\Theta} \omega(\theta)p(\theta|Y)d\theta$ for some function $\omega(\theta)$. Usually, there is a set of latent variables $u \in U \subseteq \mathbb{R}^n$ so that the augmented posterior $p(\theta, u|Y)$ is easier to analyze, yet $\int_U p(\theta, u|Y)du = p(\theta|Y)$.

In SFM, the natural candidate for such latent variables (already implicit in the notation) is the vector $u = (u_1, \ldots, u_n) \in \mathbb{R}^n_+$. Indeed, the augmented posterior in this case is simply:

$$p(\theta, u|Y) \propto \sigma_v^{-n} \prod_{i=1}^{n} \left\{ \exp\left\{ -\tfrac{1}{\sigma_v^2} \left(y_i + u_i - x_i'\beta \right)^2 \right\} p(u_i|\alpha) \right\} p(\theta). \tag{7}$$

By Y, we denote all available data like $\{y_i, x_i; i = 1, \ldots, n\}$. In alternative notation, we use y as the $n \times 1$ vector of observations on the dependent variable and X the $n \times k$ matrix containing the data on all explanatory variables. We agree that θ also contains all elements of α.

After some algebra, we can express the kernel posterior distribution as follows:

$$p(\theta,u|Y)\propto\sigma_v^{-n}\exp\left\{-\tfrac{1}{2\sigma_v^2}(y+u-X\beta)'(y+u-X\beta)\right\}\prod_{i=1}^{n}p(u_i|\alpha)p(\theta). \tag{8}$$

Given the augmented posterior $p(\theta,u|Y)$, we can apply Monte Carlo techniques[1] to explore the posterior. A standard Bayesian Monte Carlo procedure is the Gibbs sampler which proposes to alternate between obtaining random draws from $p(\theta|u,Y)$ and $p(u|\theta,Y)$. Suppose, for example, that the prior is such that

$$p(\beta,\sigma_v|\alpha)\propto p(\beta|\alpha)p(\sigma_v|\alpha), \tag{9}$$

and

$$p(\beta|\alpha)\propto\text{const.},p(\sigma_v|\alpha)\propto\sigma_v^{-(\bar{n}+1)}\exp\left(-\frac{\bar{q}}{2\sigma_v^2}\right), \tag{10}$$

where $\bar{n},\bar{q}\geq 0$ are parameters. These parameters can be elicited if we notice that from an artificial sample of size $\bar{n}$ the prior average value of σ_v^2 is $\frac{\bar{q}}{\bar{n}}$.

The prior for β is flat and the prior for σ_v is in the inverted *gamma* family. Then we can write the posterior as:

$$p(\theta,u|Y)\propto\sigma_v^{-(n+\bar{n})}\exp\left\{-\tfrac{1}{2\sigma_v^2}\left[\bar{q}+(y+u-X\beta)'(y+u-X\beta)\right]\right\}\prod_{i=1}^{n}p(u_i|\alpha)p(\theta). \tag{11}$$

The advantage of the data augmentation approach is that we can immediately extract the following conditional posterior distributions:

$$\beta|\sigma_v,\alpha,u,Y\sim N(b,V), \tag{12}$$

where $b=\left(X'X\right)^{-1}X'(y+u),V=\sigma_v^2\left(X'X\right)^{-1}$,

which follows from standard least-squares theory or the elementary Bayesian analysis of the normal linear model. Moreover, we have

[1]See Geweke (1999). For details in stochastic frontiers, see Assaf et al. (2017).

$$\frac{Q}{\sigma_v^2}|\beta,\alpha,u,Y \sim \chi^2(n+\overline{n}), \tag{13}$$

where $Q = \overline{q} + (y + u - X\beta)'(y + u - X\beta)$.

Therefore, conditionally on u we can easily generate random draws for β and σ_v. It remains to generate random draws for u and α. To generate random draws for u, we proceed element-wise and we recognize that in the posterior we have:

$$p(u_i|\beta,\sigma_v,\alpha,Y) \propto \exp\left\{-\tfrac{1}{2\sigma_v^2}(r_i+u_i)^2\right\}p(u_i|\alpha), \tag{14}$$

where $r_i = y_i - x_i'\beta$. Also the conditional posterior of α is:

$$p(\alpha|\beta,\sigma_v,u,Y) \propto \prod_{i=1}^{n} p(u_i|\alpha)p(\alpha). \tag{15}$$

In principle, both conditional posteriors are amenable to random number generation as there is a variety of techniques for drawing from arbitrary distributions. For expositional purposes, let us assume that each $u_i \sim iidN_+\left(0,\sigma_u^2\right)$ and $\alpha \equiv \sigma_u^2$. The density of the half-normal distribution is

$$p(u_i|\sigma_u) = \left(\tfrac{\pi}{2}\sigma_u^2\right)^{-1/2} \exp\left\{-\tfrac{1}{2\sigma_u^2}u_i^2\right\}. \tag{16}$$

Our prior for σ_u is:

$$p(\sigma_u) \propto \sigma_u^{-(\underline{n}+1)} \exp\left\{-\frac{\underline{q}}{2\sigma_u^2}\right\}, \tag{17}$$

where $\underline{n} \geq 0$ and $\underline{q} > 0$ are parameters of the prior. When both parameters are positive, this is a proper prior and finite integrability of the posterior, follows from Proposition 2 of Fernandez et al. (1997). If we set $\underline{q} = 0$, then we need an informative prior for the other parameters (especially β). A practical solution is to restrict each element in vector β to be uniform in an interval $[-M, M]$, where M is "very large."

Under the assumption of half-normality, the conditional posterior distribution of u_i becomes:

$$p(u_i|\beta,\sigma_v,\sigma_u,Y) \propto \exp\left\{-\tfrac{1}{2\sigma_v^2}(r_i+u_i)^2 - \tfrac{1}{2\sigma_u^2}u_i^2\right\}, \tag{18}$$

and the conditional posterior distribution of σ_u is:

$$p(\sigma_u|\beta,\sigma_v,u,Y)\propto\sigma_u^{-(n+\underline{n}+1)}\exp\left\{-\frac{\underline{q}+u'u}{2\sigma_u^2}\right\},\tag{19}$$

from which it follows that:

$$\frac{\underline{q}+u'u}{\sigma_u^2}|\beta,\sigma_v,u,Y\sim\chi^2(n+\underline{n}).\tag{20}$$

Regarding the conditional posterior distribution of u_i, we can "complete the square" to conclude:

$$u_i|\beta,\sigma_v,\sigma_u,Y\sim N_+\left(\widehat{u_i},\varphi^2\right),\tag{21}$$

where $\widehat{u_i}=-\frac{r_i\sigma_u^2}{\sigma_v^2+\sigma_u^2}$, $\varphi^2=\frac{\sigma_v^2\sigma_u^2}{\sigma_v^2+\sigma_u^2}$ and all notation has been previously introduced. Random draws from the conditional posterior of σ_u^2 can be obtained easily as

$$\sigma_u^2=\frac{\underline{q}+u'u}{z},\ \text{where}\ z\sim\chi^2(n+\underline{n}).\tag{22}$$

Drawing from the conditional posterior distribution of u_i is somewhat more involved as it involves random draws from a normal distribution truncated below at zero. In principle, one can simply draw from a normal distribution until a positive draw is obtained but this procedure is highly inefficient. A far more efficient rejection procedure has been described in Tsionas (2000).

Regarding the Gibbs sampler we should mention that given random draws from the conditional posterior distributions, say $\left\{\theta^{(s)},u^{(s)},s=1,\ldots,S\right\}=\left\{\beta^{(s)},\sigma_v^{(s)},\sigma_u^{(s)},u^{(s)},s=1,\ldots,S\right\}$ obtained S times, eventually $\left\{\theta^{(s)},s=1,\ldots,S\right\}\xrightarrow{D}p(\theta|Y)$ (Geweke 1989) so direct draws from the posterior of interest are available. From these draws, we can compute the corresponding draws for any function of interest, say $\left\{\omega\left(\theta^{(s)}\right),s=1,\ldots,S\right\}$ which may involve returns to scale, certain elasticities, etc. From these draws, one can compute easily (posterior) means and standard deviations, report histograms, or kernel density approximations to $p(\omega|Y)$, etc. For Gibbs sampling applied to stochastic frontiers, see also Koop et al. (1995).

Of course, the main function of interest is performance evaluation which is summarized by the draws $\left\{u_i^{(s)}, s = 1, \ldots, S\right\}_{i=1}^{n}$. For a specific DMU, we can collect the draws $\left\{u_i^{(s)}, s = 1, \ldots, S\right\}$ A kernel density approximation to $p(u_i|Y)$, which is the central object in Bayesian Performance Evaluation, can be easily generated using the available draws. Moreover, average inefficiency is given by:

$$E(u_i|Y) \equiv \overline{u_i} = S^{-1} \sum\nolimits_{s=1}^{S} u_i^{(s)}, \tag{23}$$

and its posterior variance is:

$$\text{var}(u_i|Y) = S^{-1} \sum\nolimits_{s=1}^{S} \left(u_i^{(s)} - \overline{u_i}\right)^2. \tag{24}$$

A 95% Bayes probability interval can be constructed easily if we sort the draws for DMU i and select the lower 2.5% and upper 97.5% values. Thus, the Bayesian approach provides a formal and straightforward way to facilitate statistical inference (which is always a problem in DEA despite recent advances). Moreover, and perhaps more importantly, all inferences about u_i are conditionally only on the data Y and, therefore, parameter uncertainty is formally taken into account.

For inferences on technical *efficiency*, we can define efficiency as $\vartheta_i = \exp(-u_i)$. Given draws $\vartheta_i^{(s)} = \exp\left\{-u_i^{(s)}\right\}, s = 1, \ldots, S$. From these draws, we can compute posterior moments and densities of ϑ_i easily.

The number of MCMC simulations, S, can be determined using Geweke's (1991) convergence diagnostic whose computation relies on testing equality of means at the start and the end of posterior simulations.

3 Model Selection

Following the Bayesian approach, model selection is straightforward. For example, conditional on normality for v_i alternative models for u_i can be estimated, for example, half-normal, exponential, Weibull, gamma, etc. Suppose we have a range of such models indexed by $\mu \in \{0, 1, \ldots, \mu^*\}$ and denote the parameter vector by θ_μ. The marginal likelihoods by definition are:

$$M_\mu(Y) = \int L(\theta_\mu; Y) p\left(\theta_\mu\right) \mathrm{d}\theta_\mu, \quad \mu \in \left\{0, 1, \ldots, \mu^*\right\}. \tag{25}$$

Suppose take model μ as a benchmark. Then we can compute the so-called Bayes factors:

$$BF_{\mu:0} = \frac{M_\mu(Y)}{M_0(Y)}, \quad \mu \in \{1, \ldots, \mu^*\}, \tag{26}$$

in favor of model μ and against model 0. It is immaterial which model is taken as the benchmark. For example, if we wish to consider the Bayes factor in favor of model μ and against model μ' we can use the identity:

$$BF_{\mu:\mu'} = \frac{M_\mu(Y)}{M_0(Y)} \cdot \frac{M_0(Y)}{M_{\mu'}(Y)} = \frac{M_\mu(Y)/M_0(Y)}{M_{\mu'}(Y)/M_0(Y)}, \quad \mu \in \{1, \ldots, \mu^*\}. \tag{27}$$

Clearly, from the *model selection* viewpoint, we ought to select the model with the highest value of the marginal likelihood. There is, however, the better alternative of *model combination* or model averaging. We can define model posterior probabilities (given that we have equal prior model probabilities) as follows:

$$\pi_\mu(Y) = \frac{M_\mu(Y)}{\sum_{\mu'=0}^{\mu^*} M_{\mu'}(Y)}, \quad \mu \in \{0, 1, \ldots, \mu^*\}. \tag{28}$$

From the viewpoint of performance evaluation suppose, we have a set of draws $\left\{u_{i,(\mu)}^{(s)}, s = 1, \ldots, S\right\}$ for a DMU i and model μ. Our "score" for performance evaluation of unit i and model μ would be technical efficiency, $\vartheta_{i,(\mu)}^{(s)} = \exp\left\{-u_{i,(\mu)}^{(s)}\right\}$. Suppose the posterior of ϑ_i from model μ is denotes by $p_{(\mu)}(\vartheta_i|Y)$. The final posterior is:

$$p(\vartheta_i|Y) = \sum_{\mu=0}^{\mu^*} \pi_\mu(Y) p\left(\vartheta_{i,(\mu)}|Y\right). \tag{29}$$

Any quantity of interest can be obtained from expressions similar to (29) which is the standard Bayesian model averaging procedure.

4 Computation of Marginal Likelihood

Computation of marginal likelihood $M(Y)$ has proved an extremely difficult issue in Bayesian analysis because the integral that defines it is intractable. One approach is to notice that we already have the identity:

$$p(\theta|Y) = \frac{L(\theta; Y)p(Y)}{M(Y)}, \tag{30}$$

known as "candidate's formula" (Chib 1995). It follows easily that

$$M(Y) = \frac{L(\theta; Y)p(Y)}{p(\theta|Y)}, \forall\, \theta \in \Theta. \tag{31}$$

As this holds for *any* parameter vector, we can choose, say, the posterior mean:

$$\overline{\theta} = E(\theta|Y) = \int_{\Theta} \theta p(\theta|Y)\mathrm{d}\theta \simeq S^{-1} \sum\nolimits_{s=1}^{S} \theta^{(s)}, \tag{32}$$

assuming $\left\{\theta^{(s)}, s = 1, \ldots, S\right\} \xrightarrow{D} p(\theta|Y)$. Therefore, we have:

$$M(Y) = \frac{L\left(\overline{\theta}; Y\right)p\left(\overline{\theta}\right)}{p\left(\overline{\theta}|Y\right)}. \tag{33}$$

The numerator is easy to compute *provided we have the likelihood in closed form*. The denominator can be approximated using a multivariate normal distribution whose mean is $\overline{\theta}$ and its covariance matrix is:

$$\overline{\sum} = \mathrm{cov}(\theta|Y) \simeq S^{-1} \sum\nolimits_{s=1}^{S} \left(\theta^{(s)} - \overline{\theta}\right)\left(\theta^{(s)} - \overline{\theta}\right)'. \tag{34}$$

See DiCiccio et al. (1997). Therefore, we can approximate the denominator by:

$$p\left(\overline{\theta}|Y\right) \simeq (2\pi)^{-d/2}|\overline{\sum}|^{-1/2}. \tag{35}$$

The approximation is quite easy and leads to the following approximation for the log marginal likelihood:

$$\log M(Y) \simeq \log L\left(\overline{\theta}; Y\right) + \log p\left(\overline{\theta}\right) + \tfrac{d}{2}\log(2\pi) + \tfrac{1}{2}\log|\overline{\sum}|. \tag{36}$$

Although the approximation is quite simple, it can be applied routinely and can be incorporated in most software, it does not account for properties of the augmented posterior $p(\theta, u|Y)$, so in a sense it ignores properties of performance. Therefore, it is not used widely.

5 Other Complications

If we wish to use the *gamma* distribution (Greene 1990; Tsionas 2000) with a known shape parameter P (say 2 or 3 in which case it is known also as the Erlang distribution), we do not have to bother with drawing from the conditional posterior of P. The fact of the matter is that several gamma distributions can be estimated with a known value of P and the results can be compared using the marginal likelihood. If, however, we insist on treating the shape parameter as unknown, with an exponential prior, say, of the form $p(P) \propto \exp(-\gamma P)$, $\gamma \geq 0$, its conditional posterior distribution does not belong to a standard family:

$$f(P) \equiv \log p(P|\beta, \sigma_v, u, Y) \doteq SP - n \log \Gamma(P), \tag{37}$$

where $S = \left(n \log \alpha - \gamma + \sum_{i=1}^{n} \log u_i\right)$.

We can try an exponential distribution as a candidate-generating density. Suppose the exponential has parameter λ, viz. $g(P; \lambda) = \lambda \exp(-\lambda P)$. The log ratio of the target to the candidate is

$$F(P, \lambda) = f(P) - \log \lambda + \lambda P = (S + \lambda)P - n \log \Gamma(P) - \log \lambda. \tag{38}$$

Using the optimal rejection procedure, the first-order conditions tell us that the optimal value of $\lambda = \lambda^*$ satisfies the nonlinear equation:

$$\overline{S} + n^{-1}\lambda - \psi\left(\lambda^{-1}\right) = 0, \tag{39}$$

where

$$\psi(x) \equiv \frac{\mathrm{d} \log \Gamma(x)}{\mathrm{d}x}, \tag{40}$$

is the *digamma* or *psi* function and $\overline{S} = n^{-1}S$. To proceed, we draw a candidate P^* from an exponential distribution with parameter λ and we accept the candidate if

$$F\left(P, \lambda^*\right) - F\left(P^*, \lambda^*\right) \geq \log U,\ U \sim \text{unif}(0, 1),\ P^* = \tfrac{1}{\lambda^*}. \tag{41}$$

The nonlinear equation is very easy to solve by bisection. Moreover, if a posterior involves an intractable normalizing constant, say

$$p(\theta|Y) \propto m(\theta)^{-N} f(\theta|Y), \tag{42}$$

where $f(\theta|Y)$ is a convenient density but the term $m(\theta)$ is complicated we can write an augmented posterior as follows:

$$p(\theta, U|Y) \propto U^{N-1} e^{-m(\theta)U} f(\theta|Y). \tag{43}$$

Using properties of the gamma distribution, we have:

$$p(\theta|Y) = \int p(\theta, U|Y) \mathrm{d}U = \propto m(\theta)^N f(\theta|Y). \tag{44}$$

The conditional posterior distribution of U is *gamma* with parameters N and $m(\theta)$. The conditional posterior of θ is $p(\theta|U, Y) = e^{-m(\theta)U} f(\theta|Y)$, for which there is a number of techniques that can be used to realize random drawings.

6 Criticisms and Replies

Bayesian performance analysis using SFMs is subject to many criticisms, most of which are, however, common to Bayesian inference in general. Some of the most common are the following:

1. The results may be sensitive to the prior.
2. It is not easy to examine alternative models and/or priors.

The first criticism is correct as long as the prior is not carefully crafted to incorporate whatever we know about the problem in hand. Since we rarely if ever we know much about efficiency in a given sector, we can, for example, adopt relatively flat priors—flat, relative to the likelihood, a fact that we can establish only after posterior analysis with a given prior has been conducted.

The second criticism is not true as alternative, flexible models can be constructed and estimated. The second criticism is not unique to Bayesian models but to SFMs as a whole.

Regarding the first criticism, we do not deny that alternative priors must be considered so that the analysis is convincing to Bayesians and non-Bayesians alike. So the issue is to examine results obtained from alternative priors, that is, to perform *posterior sensitivity analysis*.

To examine how efficiency scores change when the priors change is a more important problem. However, with modern advances in parallel computation this should not pose great problems.

7 Flexible Models

The second criticism that SFM is rigid when it comes to flexible functional forms or distributions is also unjustified. When it comes to flexible functional forms, besides of course the translog or the generalized Leontief, advances have been made in applied econometrics that allow us to consider easily such extensions. The same is true when it comes to flexible distributions for the two error components, v_i and u_i. In the sampling-theory context, for example, nonparametric models have been developed in Kumbhakar et al. (2007). Griffin and Steel (2004) and Griffin (2011) proposed Bayesian nonparametric models for stochastic frontiers, using priors that are widely used in the statistics literature.

The problem is not so much that flexible distributions can be considered at will but rather that the data may not be informative about distinguishing between the two error components when the flexible distributions demand too much information from the data, i.e., when they contain a large number of shape parameters, the two flexible distributions are in the same family, etc. Problems may also arise in relatively simple problems, for example, a normal-*gamma* frontier with an unknown shape parameter P. When P exceeds three or four, the gamma distribution becomes nearly symmetric and, therefore, distinguishing performance from the two-sided error term may be quite difficult. As a result, some care must be exercised when the model is formulated. For nonparametric, Bayesian models the reader is referred to Griffin (2011) and the references therein.

1. Given the data that we often use in empirical studies, it seems that fully nonparametric approaches do not always yield useful information and flexible semi-parametric models may be more appropriate, see, for example, Tsionas (2017) and Griffin and Steel (2004).

8 Heterogeneity

Distinguishing heterogeneity from differences in performance is important as rarely DMUs can be assumed to use the same technology.

Assuming the availability of panel data we have the model:

$$y_{it} = \alpha_i + x_{it}'\beta + v_{it} - u_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T, \qquad (45)$$

where α_is represent firm effects, x_{it} is a $k \times 1$ vector of regressors (that does not include a column of ones), β is a $k \times 1$ vector of parameters, and v_{it}, u_{it} have the same interpretation as before. The usual assumption in Bayesian treatments of SFM is that the α_is are independent of the regressors and the other two error components and

$$\alpha_i \sim iidN\left(\overline{\alpha}, \sigma_\varepsilon^2\right), \quad i = 1, \ldots, n. \tag{46}$$

Under this assumption, we can express the model in the form:

$$y_{it} = \overline{\alpha} + \varepsilon_i + x_{it}'\beta + v_{it} - u_{it}, \quad i = 1, \ldots, n, \, t = 1, \ldots, T, \tag{47}$$

where

$$\varepsilon_i \sim iidN\left(0, \sigma_\varepsilon^2\right), \quad i = 1, \ldots, n. \tag{48}$$

Define $\widetilde{x}_{it} = \left[1, x_{it}'\right]'$, $\widetilde{\beta} = \left[\overline{\alpha}, \beta'\right]'$, so that we can express the model as:

$$y_{it} = \widetilde{x}_{it}'\widetilde{\beta} + v_{it} + \varepsilon_i - u_{it}. \tag{49}$$

Collecting all observations for a given DMU we have $y_i = \left[y_{i1}, \ldots, y_{iT}\right]'$, $\widetilde{X}_i = \left[\widetilde{x}_{i1}', \ldots, \widetilde{x}_{iT}'\right]'$, and we obtain:

$$y_i = \widetilde{X}_i\widetilde{\beta} + v_i + \varepsilon_i 1_T - u_i, \quad i = 1, \ldots, n, \tag{50}$$

where 1_T is the $T \times 1$ vector of ones and v_i, u_i are $T \times 1$ vectors containing the two-sided and one-sided error terms, respectively. For simplicity we proceed under the assumptions:

$$v_{it} \sim iidN\left(0, \sigma_v^2\right), \; u_{it} \sim N_+\left(0, \sigma_u^2\right), \tag{51}$$

independently of the regressors in x_{it}, mutually independent, and independent of ε_i. Although it is possible to have the distribution $p\left(y_i|\widetilde{X}_i, \theta\right)$ in closed form and perform ML estimation, it is essential to show how the Gibbs sampler would operate in this model. First, it is useful to derive the distribution of $\xi_i = v_i + \varepsilon_i 1_T \sim N(0, \Omega)$, $\Omega = \sigma_v^2 I_T + \sigma_\varepsilon^2 J_T$, where J_T is a $T \times T$ matrix whose elements are all equal to one.

For example, it is clear that we can extract the following posterior conditional distribution of $\widetilde{\beta}$:

$$\widetilde{\beta}|\sigma_v, \sigma_u, \sigma_\varepsilon, u, Y \sim N\left(\widetilde{b}, V\right), \tag{52}$$

where Y denotes all data (y and $\widetilde{X}$), and $\widetilde{b} = \left(\widetilde{X}'\Omega^{-1}\widetilde{X}\right)^{-1}\widetilde{X}'\Omega^{-1}(y+u)$, $V = \left(\widetilde{X}'\Omega^{-1}\widetilde{X}\right)^{-1}$. The posterior conditional distribution of u_i is:

$$u_i|\widetilde{\beta}, \sigma_v, \sigma_u, \sigma_\varepsilon, Y \sim N_T^+\left(\widehat{u}_i, H\right), \tag{53}$$

where N_T^+ denotes the $T-$ variate truncated normal distribution and, after some algebra in completing the square, we obtain:

$$\widehat{u}_i = -\left(\sigma_u^2\Omega^{-1} + I_T\right)\sigma_u^2\Omega^{-1}r_i,\ r_i = y_i - \widetilde{X}_i\widetilde{\beta} + u_i, \tag{54}$$

and $H = \sigma_u^2\left(\sigma_u^2\Omega^{-1} + I_T\right)^{-1}$. It is of some interest to point out that inefficiencies are no longer independent when we use the conditioning in (53). A different result corresponding to different conditioning appears in (71) below. In fact, one can design different algorithms based on different conditioning.

The other posterior conditional distributions are in standard families. For example, the posterior conditional distribution of σ_ε is:

$$\frac{Q}{\sigma_\varepsilon^2}|\widetilde{\beta}, \sigma_v, \sigma_u, \sigma_\varepsilon, u, Y \sim \chi^2(T), \tag{55}$$

where $Q = \sum_{i=1}^n \left(y_i - \widetilde{X}_i\widetilde{\beta} + u_i\right)'\left(y_i - \widetilde{X}_i\widetilde{\beta} + u_i\right) = \sum_{i=1}^n r_i'r_i$, provided the prior is $p(\sigma_\varepsilon) \propto \sigma_\varepsilon^{-1}$. Of course, other informative priors in the inverted *gamma* family can be used.

Further, heterogeneity can be introduced if the slope coefficients are allowed to be random as in Tsionas (2006). For example, we can consider the model:

$$y_{it} = x_{it}'\beta_i + v_{it} - u_{it}, \tag{56}$$

under the convention that x_{it} includes an intercept. Here, β_i is the $k \times 1$ vector of random coefficients which we can equip with the assumption:

$$\beta_i \sim iidN_k\left(\bar{\beta}, \Sigma\right), \quad i = 1, \ldots, n. \tag{57}$$

Since we can write $\beta_i = \overline{\beta} + e_i$, $e_i \sim iidN_k(0, \Sigma)$, $\quad i = 1, \ldots, n$, we obtain:

$$y_{it} = x_{it}'\overline{\beta} + \xi_{it} - u_{it},\ \xi_{it} = v_{it} + x_{it}'e_{it}. \tag{58}$$

Clearly, $\xi_{it}|x_{it}, \sigma_v, \Sigma \sim N(0, \Phi_{it})$, where $\Phi_{it} = \sigma_v^2 + x_{it}'\Sigma x_{it}$. Collecting all observations for unit i we have:

$$y_i = X_i\beta_i + v_i - u_i, \quad i = 1, \ldots, n, \tag{59}$$

from which we have:

$$y_i = X_i\overline{\beta} + \xi_i - u_i, \quad i = 1, \ldots, n, \tag{60}$$

where $\xi_i = [\xi_{i1}, \ldots, \xi_{iT}]'$. From the assumptions of the model, we have:

$$\begin{aligned}\Phi_i = E\left(\xi_i\xi_i'\right) &= \text{diag}(\Phi_{i1}, \ldots, \Phi_{iT}) \\ &= \text{diag}\left(\sigma_v^2 + x_{i1}'\Sigma x_{i1}, \ldots, \sigma_v^2 + x_{iT}'\Sigma x_{iT}\right).\end{aligned}$$

If we collect all observations, we have:

$$y = X\overline{\beta} + \xi - u, \tag{61}$$

where $y = \left(y_1', \ldots, y_n'\right)$, $X = \begin{pmatrix} X_1 & 0 & \ldots & 0 \\ 0 & X_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & X_n \end{pmatrix}$, $\xi = \left(\xi_1', \ldots, \xi_n'\right)$ and $u = \left(u_1', \ldots, u_n'\right)$. For ξ, we have:

$$\xi \sim N_{nT}(0, \Phi),\ \Phi = \text{diag}(\Phi_1, \ldots, \Phi_n). \tag{62}$$

In this form, we can derive easily the following conditional posterior distributions:

$$\overline{\beta}|\cdot \sim N(b, V), b = \left(X'\Phi^{-1}X\right)^{-1}X'\Phi^{-1}(y + u),\ V = \left(X'\Phi^{-1}X\right)^{-1}. \tag{63}$$

If we are nor interested in estimating the different slope coefficients but only their average, then (63) provides all we need. When, however the different slope coefficients are of interest as well, then it is best to proceed as follows.

To draw the random coefficients, we focus on the original representation of the model:

$$y_i = X_i\beta_i + v_i - u_i, \quad i = 1, \ldots, n. \tag{64}$$

Along with the "prior": $\beta_i \sim iidN_k\left(\overline{\beta}, \Sigma\right), \quad i = 1, \ldots, n$, completing the square yields the following expression for the conditional posterior distribution of the random coefficients:

$$\begin{aligned} &\beta_i|\cdot \sim N_k(b_i, V_i), \\ &b_i = \left(X_i'X_i + \sigma_v^2\Sigma^{-1}\right)^{-1} X_i'(y_i + u_i), \\ &V_i = \sigma_v^2\left(X_i'X_i + \sigma_v^2\Sigma^{-1}\right)^{-1}. \end{aligned} \tag{65}$$

If the prior of σ_v is $p(\sigma_v) \propto \sigma_v^{-(\overline{n}+1)} \exp\left\{-\frac{\overline{q}}{2\sigma_v^2}\right\}$, $\overline{n}, \overline{q} \geq 0$, then its conditional posterior distribution is:

$$\frac{\sum_{i,t}\left(y_{it} + u_{it} - X_{it}'\beta_i\right)^2}{\sigma_v^2}|\cdot \sim \chi^2(nT). \tag{66}$$

The conditional posterior for $\overline{\beta}$ has a particularly convenient form:

$$\begin{aligned} \overline{\beta}|\cdot &\sim N_k(b, V), \\ b = n^{-1}\sum_{i=1}^n \beta_i,\ V &= n^{-1}\Sigma, \end{aligned} \tag{67}$$

although an alternative form has been provided above. The conditional posterior for Σ is:

$$\begin{aligned} p(\Sigma|\cdot) &\propto |\Sigma|^{-n/2} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^n \left(\beta_i - \overline{\beta}\right)'\Sigma^{-1}\left(\beta_i - \overline{\beta}\right)\right\} p(\Sigma) \\ &= |\Sigma|^{-n/2} \exp\left\{-\tfrac{1}{2} trA\Sigma^{-1}\right\} p(\Sigma), \end{aligned} \tag{68}$$

where $A = \sum_{i=1}^n \left(\beta_i - \overline{\beta}\right)\left(\beta_i - \overline{\beta}\right)'$ and $p(\Sigma)$ is the prior of the different elements of Σ. The form of the conditional likelihood suggests that a prior for Σ can be defined as:

$$p(\Sigma) \propto |\Sigma|^{-(\overline{n}+k+1)/2} \exp\left(-\tfrac{1}{2} tr\overline{A}\Sigma^{-1}\right), \tag{69}$$

where $\overline{n} \geq 0$ and $\overline{A}$ is a $k \times k$ positive semi-definite matrix. Then the conditional posterior for Σ is:

$$p(\Sigma|\cdot) \propto |\Sigma|^{-(n+\overline{n}+k)/2} \exp\left\{-\tfrac{1}{2} tr\left(A + \overline{A}\right)\Sigma^{-1}\right\}, \tag{70}$$

which is in the Wishart family.

Regarding the conditional posterior for u_{it} we have a standard result:

$$\begin{aligned} u_{it} &\sim N_{+}\left(\hat{u}_{it}, h^2\right), \\ \hat{u}_{it} &= -\frac{\sigma_u^2\left(y_{it} - x_{it}'\beta_i\right)}{\sigma_v^2 + \sigma_u^2}, h^2 = \frac{\sigma_v^2\sigma_u^2}{\sigma_v^2 + \sigma_u^2}. \end{aligned} \tag{71}$$

An alternative expression can be derived from $y = X\overline{\beta} + \xi - u$, where we draw jointly all inefficiency measures:

$$\begin{aligned} &u \sim N_{+}^{nT}\left(\widehat{u}, V\right), \\ &\hat{u} = -\sigma_u^2\left(\sigma_u^2\Phi^{-1} + I_{nT}\right)^{-1}\Phi^{-1}\left(y - X\overline{\beta}\right), V = \sigma_u^2\left(\sigma_u^2\Phi^{-1} + I_{nT}\right)^{-1}. \end{aligned} \tag{72}$$

It is beyond our scope in this chapter to propose efficient rejection procedures for this multivariate density.

9 Bayesian Fixed-Effect Models

Suppose we consider the model:

$$y_{it} = \alpha_i + x_{it}'\beta + v_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T, \tag{73}$$

under the assumption that the individual effects (α_i) contain only technical inefficiency. The standard corrected least squares estimator (COLS) provides a relative inefficiency measure as:

$$u_i^* = -\left(\widehat{\alpha}_i - \max_{j=1,\ldots,n}\widehat{\alpha}_j\right), \tag{74}$$

where the $\widehat{\alpha}_j$ are obtained through the LSDV estimator. By construction, we have $u_i^* \geq 0$. Of course, a certain DMU will always be fully efficient. Let us call, this DMU the "benchmark." The classical estimator in this context is the CSS estimator (Cornwell et al. 1990). The model relies on a flexible parametrization of inefficiency in the context of the model:

$$y_{it} = x_{it}'\beta + a_{0,it} + a_{1,it}t + a_{2,it}t^2 + v_{it}, \tag{75}$$

where v_{it} is a usual stochastic error, and $a_{0,it}, a_{1,it}, a_{2,it}$ are firm-specific coefficients. The flexible trend function captures technical inefficiency after a transformation similar to COLS.

A Bayesian equivalent has been proposed that relies on the following. For simplicity, we can assume: $v_{it} \sim iidN\left(0, \sigma_v^2\right)$.

$$\widetilde{u}_i^{(s)} = -\left(\alpha_i^{(s)} - \max_{j=1,\ldots,n} \alpha_j^{(s)}\right), \quad s = 1, \ldots, S, \tag{76}$$

where $\left\{\alpha_i^{(s)}, s = 1, \ldots, S\right\}_{i=1}^{n}$ denotes the set of Gibbs sampling draws for α_is. Finally, as in Koop et al. (1997), the inefficiency measure is

$$\widetilde{u}_i = S^{-1} \sum\nolimits_{s=1}^{S} \widetilde{u}_i^{(s)}, \quad i = 1, \ldots, n, \tag{77}$$

viz. an average across MCMC draws. The advantage of the estimator is that, *in general*, all inefficiency estimates will be non-zero, and the "benchmark" is not, in general, a *given* DMU. In fact, one can compute the probability that a certain DMU is fully efficient—by counting the number of times it has $\widetilde{u}_i^{(s)} = 0$ and dividing by S. Whether or not this estimator performs any better compared to CSS is an open question.

This Bayesian "fixed effects estimator" can be extended to allow for heterogeneity in β. As a matter of fact, however, the distinction between fixed and random coefficients in Bayesian analysis is elusive as, eventually, all coefficients are random variables. The generic disadvantage of both the Bayesian and the sampling-theory paradigm, in this context, is that inefficiency is assumed to be time-invariant and it is identified with individual effects. Therefore, we have to consider this problem as in the following section.

10 Heterogeneity and the Treatment of Individual Effects

In the true fixed-effect model, we assume that:

$$y_{it} = \alpha_i + x_{it}'\beta + v_{it} - u_{it}, \quad i = 1, \ldots, n, \, t = 1, \ldots, T, \tag{78}$$

where $v_{it} \sim N\left(0, \sigma_v^2\right)$, $u_{it} \sim N_+\left(0, \sigma_u^2\right)$, which are mutually independent and independent of the regressors. The α_is are fixed effects. The model can be estimated using ML and Greene (2005) has provided an *ingenious way*

to construct a Gauss-Newton iteration even when n is large. This construction applies to more general nonlinear model with fixed effects.

From the Bayesian point of view, we have already provided the computational details to perform statistical analysis in this model and, in fact, a more general model with DMU-specific coefficients, β_i.

Individual effects can be distinguished from persistent technical inefficiency, by using the following model (Tsionas and Kumbhakar 2012):

$$y_{it} = \alpha_i + \lambda_i^+ + x_{it}'\beta + v_{it} - u_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T, \tag{79}$$

where $\lambda_i^+ \geq 0$ denotes persistent technical inefficiency. One standard assumption would be that[2]:

$$\lambda_i^+ \sim iidN_+\left(0, \sigma_\lambda^2\right), \quad i = 1, \ldots, n. \tag{80}$$

Bayesian analysis using the Gibbs sampler is a simple extension of what we described in previous sections of this chapter. A more general model is to allow for time effects as well:

$$y_{it} = \alpha_i + \mu_t + \lambda_i^+ + \tau_t^+ + x_{it}'\beta + v_{it} - u_{it}, \; i = 1, \ldots, n, t = 1, \ldots, T, \tag{81}$$

where μ_t is a general time effect, and $\tau_t^+ \geq 0$ is a DMU-specific inefficiency component, for which a standard distribution assumption would be:

$$\tau_t^+ \sim iidN_+\left(0, \sigma_\tau^2\right), \quad t = 1, \ldots, T. \tag{82}$$

Then, $u_{it} \geq 0$ can be interpreted as the transitory component of inefficiency. Estimating permanent and transitory components is of great interest in performance analysis. From the point of view of policy analysis, the objective is to see how the sectoral component τ_t^+ has evolved over time and, naturally, to examine how it can be reduced.

As both permanent and transitory inefficiency components do not depend on contextual or environmental variables, as they are commonly called, an important issue remains unaddressed. This is what we take up in the next section.

[2]All random effects are mutually independent and independent of the regressors for all DMUs and at all time periods.

11 Inefficiency Determinants

In the standard model:

$$y_{it} = x_{it}'\beta + v_{it} - u_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T, \tag{83}$$

one way to provide inefficiency determinants is to assume:

$$u_{it}|z_{it}, \gamma, \sigma_u \sim N_+\left(z_{it}'\gamma, \sigma_u^2\right), \tag{84}$$

where z_{it} is an $m \times 1$ vector of contextual variables and γ is a vector of coefficients. The model has been introduced by Battese and Coelli and others. It is interesting to see how the likelihood/posterior change as a result of this assumption. The density of u_{it} is:

$$p(u_{it}|z_{it}, \gamma, \sigma_u) = \frac{1}{\sqrt{2\pi}\sigma_u^2\Phi\left(z_{it}'\gamma/\sigma_u\right)} \exp\left\{-\frac{1}{2\sigma_u^2}\left(u_{it} - z_{it}'\gamma\right)^2\right\}, \tag{85}$$

where $\Phi(\cdot)$ is the standard normal distribution function. This form suggests the reparametrization $\delta = \gamma/\sigma_u$ from which we obtain:

$$p(u_{it}|z_{it}, \gamma, \sigma_u) = \frac{1}{\sqrt{2\pi}\sigma_u^2\Phi\left(z_{it}'\delta\right)} \exp\left\{-\frac{1}{2}\left(\sigma_u^{-1}u_{it} - z_{it}'\delta\right)^2\right\}. \tag{86}$$

The reparametrization has been suggested by Greene (chapter "Micro Foundations of Earnings Differences" in Fried et al. 1993) in order to stabilize ML procedures. Assuming $v_{it} \sim iidN\left(0, \sigma_v^2\right)$ independently of $\{u_{i\tau}, x_{i\tau}, z_{i\tau}; \tau = 1, \ldots, T\}$ it can be shown, after some algebra that the conditional posterior distribution of the one-sided component is:

$$\begin{aligned} &u_{it}|\cdot \sim N_+\left(\widehat{u}_{it}, h^2\right), \\ &\widehat{u}_{it} = \frac{-\sigma_u^2(y_{it} - x_{it}')\beta + \sigma_v^2 z_{it}'\gamma}{\sigma_v^2 + \sigma_u^2}, \; h^2 = \frac{\sigma_v^2\sigma_u^2}{\sigma_v^2 + \sigma_u^2}. \end{aligned} \tag{87}$$

From the form of the conditional posterior distribution which can also be used in sampling-theory analysis, provided estimates are available, we see clearly the effect of contextual variables on inefficiency, see also Assaf et al. (2017).

Although the posterior conditional distributions of β, σ_v, u are in standard families this is not so for δ and σ_u but efficient computational procedures can be devised to draw random numbers from the respective conditional posterior distributions.

There are alternative ways to incorporate environmental variables. Given the standard model, a more reasonable formulation would be

$$\log u_{it} = z'_{it}\gamma + \varepsilon_{it}, \varepsilon_{it} \sim iidN(0, \sigma_\varepsilon^2), \tag{88}$$

which allows for a lognormal distribution of inefficiency. Alternatively, if inefficiency follows an exponential distribution, say $p(u_{it}) = \theta_{it}^{-1} \exp\left(-\theta_{it}^{-1} u_{it}\right)$, $u_{it} \geq 0$, since the expected value is $E(u_{it}) = \theta_{it}$ we can adopt a model of the form:

$$\log \theta_{it} = z'_{it}\gamma \text{ or } \theta_{it} = \exp\left(z'_{it}\gamma\right). \tag{89}$$

Using Bayesian techniques, it is also possible to analyze models with unobserved heterogeneity:

$$\log \theta_{it} = z'_{it}\gamma + \varepsilon_{it},\ \varepsilon_{it} \sim iidN(0, \sigma_\varepsilon^2). \tag{90}$$

Empirically, it seems appropriate to modify such models to allow for dynamics, for example:

$$\log u_{it} = z'_{it}\gamma + \rho \log u_{i,t-1} + \varepsilon_{it},\ \varepsilon_{it} \sim iidN(0, \sigma_\varepsilon^2), \tag{91}$$

where u_{i0} is given or $\log u_{i0}$ follows a certain distribution depending on unknown parameters. For example, if a steady state exists, we can assume

$$\log u_{i0} = z'_{i0}\tfrac{\gamma}{1-\rho} + e_{it},\ e_{it} \sim iidN\left(0, \tfrac{\sigma_\varepsilon^2}{1-\rho^2}\right). \tag{92}$$

More generally, we can assume:

$$\log u_{i0} \sim N\left(\underline{a}_{i0}, \underline{\omega}^2\right), \quad i = 1, \ldots, n, \tag{93}$$

where $\underline{u}_{i0}, \underline{\omega}^2$ are unknown parameters. In a hierarchical fashion, it is possible to assume:

$$\underline{a}_{i0} \sim iidN\left(\underline{a}, \underline{b}^2\right), \quad i = 1, \ldots, n, \tag{94}$$

where the hyperparameters $\underline{a}, \underline{b}^2$ are, usually, fixed. Other attempts have been made where in the standard model it is assumed that:

$$u_{it} = \exp\left(z_{it}'\gamma\right)u_i^+, \quad i = 1, \ldots, n, \tag{95}$$

and $u_i^+ \sim iidN_+\left(0, \sigma_u^2\right)$. Such models usually produce formulations that are less computationally demanding.

12 Endogeneity

Endogeneity has been neglected for a long time but attempts have been made to take it into account. In the standard model:

$$y_{it} = x_{it}'\beta + v_{it} - u_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T, \tag{96}$$

it is usual that the regressors are endogenous, that is correlated with v_{it} or even u_{it} or both. There may be economic reasons why this is so but we will come to this matter in the next section. In such cases, we need instruments $z_{it} \in \mathbb{R}^m$ which can be used to obtain a reduced form:

$$x_{it} = \Pi z_{it} + V_{it,*}, \tag{97}$$

where Π is a $k \times m$ matrix of coefficients, and

$$V_{it} = \left[v_{it}, V_{it,*}'\right]' \sim N_{k+1}(0, \Sigma). \tag{98}$$

We assume first that u_{it} is independent of all other error components, the regressors and the instruments. Joint estimation by ML or Bayesian methods is not difficult in this formulation. If the regressors and inefficiency are not independent, one way to deal with the issue is to incorporate the regressors in u_{it} as "environmental variables" as in the previous section. It is possible, for example, to change the reduced form as follows:

$$\begin{bmatrix} x_{it} \\ \log u_{it} \end{bmatrix} = \Pi z_{it} + V_{it,*}, \tag{99}$$

with the apparent modifications in the dimensions of Π and $V_{it,*}$. More generally, the reduced form may not be linear in which case we can consider formulations like:

$$\begin{bmatrix} x_{it} \\ \log u_{it} \end{bmatrix} = \Pi(z_{it}) + V_{it,*}, \tag{100}$$

where $\Pi()$ is a matrix function. It is obvious that lagged values of x_{it} can be used as instruments, although in this case the issue of weak instruments arises and proper tests can be developed. In the Bayesian literature, such tests are lacking and in the sampling-theory literature this is an active area of research. General investigations show that weak instruments are likely to produce irregularly shaped posteriors in linear models. For this limited information approach to endogeneity, see Kutlu (2010) and Tran and Tsionas (2013, 2015, 2016).

13 Economic Reasons for Endogeneity

Let us recall the cost minimization problem:

$$\min_{x \in \mathbb{R}_+^k} : w'x,\ T(x,y) \leq 1, \tag{101}$$

and

$$\max_{x \in \mathbb{R}_+^K, y \in \mathbb{R}_+^M} : p'y - w'x,\ T(x,y) \leq 1, \tag{102}$$

where $T(x,y) \leq 1$ describes the technology x, y represent inputs and outputs, and w, p denote input and output prices, respectively. There are many theoretically equivalent ways to describe the technology, like input-oriented distance functions (IDF) and output-oriented distance functions (ODF). Cost minimization suggests that inputs are endogenous to the firm, while output plus input prices are predetermined. Profit maximization suggests that given input and output prices both inputs and outputs are endogenous. Output endogeneity and input exogeneity would result from a problem of revenue maximization. Of course, different assumptions may be reasonable or more appropriate in different contexts.

Output distance functions are homogeneous of degree one in outputs while input distance functions are homogeneous of degree one in inputs. Therefore, for an ODF we can write (in log terms):

$$y_1 = F\left(\tilde{y}_2, \ldots, \tilde{y}_M, x\right) + v - u = F(\tilde{y}, x) + v - u, \tag{103}$$

where $\widetilde{y}_m = \log \frac{Y_m}{Y_1}$, $m = 2, \ldots, M$, outputs in levels are $Y_1, \ldots, Y_M$, inputs are $X_1, \ldots, X_K$ and $u \geq 0$ represents technical inefficiency. The literature has ignored for a while the fact that the $\widetilde{y}_m$s are endogenous under revenue or profit maximization. In the second case, the xs will be endogenous as well. Similarly, for the IDF we have:

$$x_1 = F\left(y, \widetilde{x}_2, \ldots, \widetilde{x}_K\right) + v - u = F(y, \widetilde{x}) + v + u, \tag{104}$$

where $\widetilde{x}_k = \log \frac{X_k}{X_1}$, $k = 2, \ldots, K$. The $\widetilde{y}_m$s or $\widetilde{x}_m$s that are endogenous need to be handled using the reduced form approach above, or perhaps better, incorporate the first-order conditions resulting from cost minimization or profit maximization.

The problem was long-standing for some time as the IDF or ODF provides one equation, and therefore, we need an additional $M - 1$ or $K - 1$. Fernandez et al. (2002) have proposed a certain solution of the problem by providing additional equations that are not, in essence, different from a reduced form approach.

In principle, an ODF and an IDF are both representations of the same technology so the choice between them does not appear to be an issue of substance. However, it seems that an IDF is convenient under cost minimization where inputs and endogenous and outputs are predetermined for the firm. The ODF is more convenient under revenue maximization. In practice, however, it may turn out that when we use the IDF, outputs may be statistically endogenous. Therefore, the economic and econometric assumptions may be in conflict. The choice between estimating an IDF or and ODF is difficult and has not been taken up in the literature in a satisfactory way, to the best of the author's knowledge.

14 Greene's Problem: Estimation of Technical and Allocative Inefficiency

Known as Greene's problem the joint estimation of technical and allocative inefficiency has received great interest in the literature. Kumbhakar (1997) has proposed a model where technical and allocative inefficiency can be jointly considered and Kumbhakar and Tsionas (2005a, b) have taken up Bayesian inference in the model in the panel data and cross-sectional data setting. The model starts from cost minimization assuming that firms have misperceived input prices as $w_j^* = w_j e^{-\xi_j}$ where ξ_k can be interpreted as

allocative distortion. The model in Kumbhakar (1997) finally produces the following equations[3]:

$$\log C_{it}^{a} = \log C_{it}^{*} + \log G_{it} + v_{it} + u_i, \tag{105}$$

$$S_{j,it}^{a} = S_{j,it}^{0} + \eta_{j,it},\ j = 1, \ldots, K, \tag{106}$$

where C_{it}^{a} and $S_{j,it}^{a}$ represent actual (observed) cost and shares, $S_{j,it}^{o}$ is the usual expression from a translog cost function, and G_{it} and η_{it} are complicated expressions depending on the ξ_js. More specifically, we have the following system of equations:

$$\begin{aligned}\log C_{it}^{0} = {} & \alpha_0 + \sum_{j=1}^{K} \alpha_j \log w_{j,it} + \gamma_y \log y_{it} + \tfrac{1}{2}\gamma_{yy}(\log y_{it})^2 \\ & + \tfrac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K} \beta_{jk} \log w_{j,it} \log w_{k,it} + \sum_{j=1}^{K} \gamma_{jy} \log w_{j,it} \log y_{it} \\ & + \alpha_t t + \tfrac{1}{2}\alpha_{tt} t^2 + \beta_{yt} \log y_{it} t + \sum_{j=1}^{K} \beta_{jt} \log w_{j,it} t,\end{aligned} \tag{107}$$

$$S_{j,it}^{0} = \alpha_j + \sum_{k=1}^{K} \beta_{jk} \log w_{k,it} + \gamma_{jy} \log y_{it} + \beta_{jt} t, \tag{108}$$

$$\begin{aligned}\log C_{it}^{AL} = {} & \log G_{it} + \sum_{j=1}^{K} \alpha_j \xi_{j,i} + \sum_{j=1}^{K}\sum_{k=1}^{K} \beta_{jk} \xi_{j,i} \log w_{k,it} \\ & + \tfrac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K} \beta_{jk} \xi_{j,i} \xi_{j,k} + \sum_{j=1}^{K} \gamma_{jy} \xi_{j,i} + \sum_{j=1}^{K} \beta_{jt} \xi_{j,i} t,\end{aligned} \tag{109}$$

where

$$G_{it} = \sum_{j=1}^{K} S_{j,it}^{*} e^{-\xi_{j,i}}, \tag{110}$$

$$S_{j,it}^{*} = \alpha_j + \sum_{k=1}^{K} \beta_{jk} \log w_{k,it}^{*} + \gamma_{jy} \log y_{it} + \beta_{jt} t = S_{j,it}^{0} + \sum_{k=1}^{K} \beta_{jk} \xi_k. \tag{111}$$

Moreover, we have:

$$\eta_{j,it} = \frac{S_{j,it}^{0}\left\{1 - G_{it} e^{\xi_{j,i}}\right\} + \sum_{k=1}^{K} \beta_{jk} \xi_k}{G_{it} e^{\xi_{j,i}}}. \tag{112}$$

[3]We assume for simplicity that we have only one output and that technical inefficiency and allocative distortion parameters ξ are time-invariant.

The notation $\xi_{j,i}$ provides the jth allocative distortion parameter for firm i. Therefore, the $\eta_{j,it}$ are the deviations of the actual cost shares from their optimum values, and they are nonlinear functions of allocative inefficiencies $\xi_2, \ldots, \xi_K$, and, of course, the data. In obvious notation, we can write the system as follows:

$$\begin{aligned} y_1 &= X_1(\xi)\beta_1 + \log G(\xi, \beta) + v_1 + u \otimes 1_T, \\ y_j &= X_2\beta_j + \eta_{j-1}(\xi, \beta) + v_j, j = 2, \ldots, K, \end{aligned} \tag{113}$$

where y_1 denotes the vector of all observations on log costs, y_j denotes the jth ($j = 2, \ldots, K$) share. X_1 and X_2 denote the matrices of observations of the cost function and the share equations, respectively. We assume that $v = \left[v_1', \ldots, v_K'\right]' \sim N_{KnT}(0, \Sigma)$ supposing we have n firms and T time observations. Kumbhakar and Tsionas (2005a) assume that $\xi \sim N_{n(K-1)}(0, \Omega \otimes I_n)$ so that allocative distortions are random variables.

Statistical inference in the system is complicated by the fact that it is highly nonlinear in ξ, and therefore, specialized MCMC methods are devised. One particular feature of the system is that it is linear in parameters β conditional on ξs. Estimation of technical inefficiency can be performed using the posterior expectation of u while estimation of allocative inefficiency can be performed using the posterior expectation of $\log C_{it}^{AL}$. The same problem can be treated when only cross-sectional data is available, an issue that is taken up in Kumbhakar and Tsionas (2005b).

It would be interesting to drop the assumption that allocative distortions are random variables and assume, instead, that they are fixed parameters. Greene's (2005) technique could have been employed to reduce considerably the burden of computation as the required derivatives are provided in Kumbhakar and Tsionas (2005a). In this way, it would not be necessary to assume that technical inefficiency and allocative distortions are statistically independent. However, we should remark that a posteriori they are not statistically independent.

15 Fitting General Stochastic Frontier Models

For a general stochastic cost frontier model of the form $y_{it} = x_{it}'\beta + v_{it} + u_{it}$ under independence of the two error components, a certain problem arises when the distribution of v_{it} is not normal and the distribution of u_{it} is anything other than half-normal or exponential. Tsionas (2012) notices that if we have the characteristic functions $\varphi_v(\tau)$ and $\varphi_u(\tau)$

($\tau \in \mathbb{R}$) of the two error components then the characteristic function of the composed error $\varepsilon_{it} = v_{it} + u_{it}$ is simply: $\varphi_\varepsilon(\tau) = \varphi_v(\tau)\varphi_u(\tau)$.

The characteristic function can, in turn, be inverted, using the Fourier transform and yield directly the density of the composed error. The required formula is $f_\varepsilon(\varepsilon) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-\iota\tau\varepsilon}\varphi_\varepsilon(\tau)d\tau$ where $\iota = \sqrt{-1}$. The inversion can be implemented using the Fast Fourier Transform (FFT) which is widely available in software.

The density is evaluated at specific points and to obtain the density at the observed points, we must use interpolation which is, usually, quite accurate.

Tsionas (2012) considered models "with output-oriented stochastic frontier models whose one-sided errors have a distribution other than the standard ones (exponential or half-normal). The *gamma* and *beta* distributions are leading examples. *Second*, with input-oriented stochastic frontier models which are common in theoretical discussions but not in econometric applications. *Third*, with two-tiered stochastic frontier models when the one-sided error components follow *gamma* distributions. *Fourth*, with latent class models with *gamma* distributed one-sided error terms. *Fifth*, with models whose two-sided error component is distributed as stable Paretian and the one-sided error is *gamma*. The principal aim is to propose approximations to the density of the composed error based on the inversion of the characteristic function (which turns out to be manageable) using the Fourier transform. Procedures that are based on the asymptotic normal form of the log-likelihood function and have arbitrary degrees of asymptotic efficiency are also proposed, implemented and evaluated in connection with output-oriented stochastic frontiers. The new methods are illustrated using data for US commercial banks, electric utilities, and a sample from the National Youth Longitudinal Survey."

16 More on Endogeneity Issues

SFM of the form $y_{it} = x_{it}'\beta + v_{it} + u_{it}$ usually ignore the fact that x_{it}s may be endogenous, that is correlated with the error term v_{it} and/or u_{it}. In Sect. 17, we have summarized some recent limited information approaches to deal with this issue. An additional concern is, however, that such approaches are not always compatible with the economics of the problem. Suppose for simplicity, we have a deterministic production function: $Y = e^{\beta_0}\prod_{k=1}^{K} X_k^{\beta_k}$. If we use lower-case letters to denote logs and estimate the Cobb-Douglas functional form:

$$y = \beta_0 + \sum_{k=1}^{K} \beta_k x_k + v, \tag{114}$$

it is well known since Marschak and Andrews (1944) that there are endogeneity problems under certain behavioral assumptions. For example, under the widely used cost-minimization assumption, we obtain the first-order conditions:

$$\begin{aligned} y &= \beta_0 + \sum_{k=1}^{K} \beta_k x_k + v_1, \\ \log \frac{\beta_k}{\beta_1} &= w_k - w_1 + x_k - x_1 + v_k,\ k = 2, \ldots, K, \end{aligned} \tag{115}$$

where the endogenous variables are $x_k, k = 1, \ldots, K$, output is considered exogenous,$v_1, \ldots, v_k$ are error terms, and $w_1, \ldots, w_k$ denote logs of input prices. The above is a simple simultaneous equations model, nonlinear in the parameters which can be estimated by FIML *provided we have input relative prices*.

Very often we lack this luxury. However, the system motivates us to use it in this form and make assumptions about input relative prices. For example, with panel data, the system above is:

$$\begin{aligned} y_{it} &= \beta_0 + \sum_{k=1}^{K} \beta_k x_{k,it} + v_{1,it}, \\ \log \frac{\beta_k}{\beta_1} &= \delta_{k,it} + x_{k,it} - x_{1,it} + v_{k,it}, \quad k = 2, \ldots, K, i = 1, \ldots, n, \end{aligned} \tag{116}$$

where $\delta_{k,it} = w_{k,it} - w_{1,it}$. One assumption that can be used is that

$$\delta_{k,it} = \lambda_{k,i} + \mu_{k,t},\ k = 2, \ldots, K, \tag{117}$$

where $\lambda_{k,i}, \mu_{k,t}$ are input-specific firm and time effects.[4] If one believes that prices are approximately the same for all firms in a given time period, then one can simply set $\lambda_{k,i} = \overline{\lambda}_k$, a constant to be estimated.

Of course many other assumptions are possible and the restriction to a Cobb-Douglas functional form is inessential. If we denote

$$\delta_{it} = \left[\delta_{it,2}, \ldots, \delta_{it,K}\right], \tag{118}$$

it is possible to assume a vector autoregressive scheme (VAR) of the form:

[4]These can be identified even if there are firm and time effects in the production function.

$$\delta_{it} = \alpha + B\delta_{i,t-1} + \Gamma z_{it} + e_{it},\ e_{it} \sim iidN(0, \Sigma), \tag{119}$$

where z_{it} is a vector of certain predetermined variables like other price indices and related variables. With the VAR scheme, of course, the model becomes cumbersome and special Bayesian filtering techniques are required to explore the posterior distribution.

These technical points aside, the fact of the matter is that *the first-order conditions from certain behavioral assumptions do provide additional equations to allow for endogenization of the regressors even when prices are not available*. This point has been largely overlooked in the literature. The approach has been followed by Atkinson and Tsionas (2016). As their data contain input and output price data, they append price equations (where prices are related to marginal products) for inputs to the directional distance function to obtain a cost-minimization directional distance system and the price equations for all good inputs and outputs to obtain a profit-maximization directional distance system. They identify the directions for bad inputs and bad outputs, which lack prices, using methods explained below.

Using MCMC methods they estimate these systems, obtaining estimates of all structural parameters, optimal directions, measures of technical efficiency, productivity growth, technical change, and efficiency change, and estimates of the implied optimal percent changes in inputs and outputs. These directions are those that would prevail in the industry if firms were cost minimizers or profit maximizers. That is, they are estimating directional distance functions, not with directions chosen a priori, but with optimal directions chosen that are consistent with cost minimization or profit maximization.

It seems that using information from first-order conditions is a plausible way of dealing with endogeneity issues. Even when information on relative prices is unavailable, the researcher may assume that the missing prices are latent and follow a particular model. An alternative is to use individual-specific and time-specific effects for the missing prices.

Endogeneity arises also as a major problem in estimating productivity (Olley and Pakes 1996; Levinsohn and Petrin 2003) so these principles can be applied in this context as well. One such approach is taken up in Gandhi et al. (2013).

17 A Lognormal Inefficiency Effects Models

Models with inefficiency or environmental effects have been quite popular in the efficiency literature. The workhorse of inefficiency effects model is the so-called Battese and Coelli model:

$$\begin{aligned} &y_{it} = x_{it}'\beta + v_{it} + u_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T, \\ &v_{it} \sim iidN(0, \sigma_v^2), u_{it} \sim N_+(z_{it}'\gamma, \sigma_u^2), \end{aligned} \tag{120}$$

where x_{it} is an $k \times 1$ vector of regressors, z_{it} is an $m \times 1$ vector of environmental variables and β, γ are $k \times 1$ and $m \times 1$ vector of parameters. In this section, we propose an alternative inefficiency effects model:

$$\log u_{it} \sim N\left(z_{it}'\gamma, \sigma_u^2\right). \tag{121}$$

Bayesian analysis of the model is quite standard given the tools that we have described in this chapter. Specifically, applying a Gibbs sampler is straightforward as drawing random numbers from the posterior conditional distributions of $\beta, \gamma, \sigma_v^2$ and σ_u^2 is not difficult. The cumbersome part is to draw from the posterior conditional distributions of inefficiencies. These posterior conditional distributions have the following form:

$$\begin{aligned} &p(u_{it}|\beta, \gamma, \sigma_v, \sigma_u, y, X) \propto \\ &\exp\left\{-\tfrac{1}{2\sigma_v^2}(R_{it} - u_{it})^2 - \tfrac{1}{2\sigma_u^2}\left(\log u_{it} - z_{it}'\gamma\right)^2 - \log u_{it}\right\}, u_{it} > 0, \end{aligned} \tag{122}$$

where $R_{it} = y_{it} - x_{it}'\beta$. If we reparametrize to $h_{it} = \log u_{it}$ we have:

$$\begin{aligned} &p(h_{it}|\beta, \gamma, \sigma_v, \sigma_u, y, X) \propto \\ &\exp\left\{-\tfrac{1}{2\sigma_v^2}(R_{it} - e^{h_{it}})^2 - \tfrac{1}{2\sigma_u^2}\left(h_{it} - z_{it}'\gamma\right)^2 + h_{it}\right\}, u_{it} > 0. \end{aligned} \tag{123}$$

To draw from this posterior conditional distribution, we can draw $h_{it} \sim N(z_{it}'\gamma, \sigma_u^2)$ and accept the draw if $\exp\left\{-\frac{1}{2\sigma_v^2}(R_{it} - e^{h_{it}})^2\right\} \geq U$, where U is a standard uniform random number. This procedure is not, in general, efficient as we draw from the "prior" and we accept based on the likelihood. A more efficient alternative is not immediately obvious, unless $R_{it} - 2e^{h_{it}} < 0$ in which case (123) is log-concave, and specialized algorithms can be used.

We apply the new model to artificial data and Greene's electricity data. For the artificial data, we use $k = m = 3$. Matrices X and Z contain a

Table 1 Empirical results for artificial data

	Posterior mean	Posterior s.d.
β_1	1.0014	0.0078
β_2	1.0027	0.0047
β_3	1.0039	0.0048
γ_1	−3.0002	0.0086
γ_2	−0.1001	0.0083
γ_3	−0.0993	0.0088
σ_v	0.1032	0.0060
σ_u	0.1925	0.0231
Inefficiency	0.051	0.007
Efficiency	0.9503	0.0066

column of ones and the remaining columns are random numbers generated from a standard normal distribution. We set the element of β to 1. All elements of γ are set to 0.1 except the intercept which is -3. Moreover, $\sigma_v = 0.1$ and $\sigma_u = 0.2$. We use 15,000 iterations of MCMC the first 5000 of which are discarded to mitigate start-up effects. Priors for β and γ are flat. The prior for σ_v and σ_u is $p(\sigma) \propto \sigma^{-(\nu+1)} e^{-b/(2\sigma^2)}$ where $\nu = 1$ and $b = 0.001$. The results are reported in Table 1.

Next, we turn attention to Greene's electricity data[5] which contains cost data for a cross section of 145 electric US utilities. We estimate a cost function of the form

$$\log(C/p_L) = F\left(\log \frac{p_K}{p_L}, \log \frac{p_F}{p_L}, \log y; \beta\right) + v + u,\ u \geq 0, \quad (124)$$

where p_L, p_K, p_F are prices of labor, capital, and fuel, y is output and C is cost. Here, F is the translog functional form. As determinants of inefficiency, we have the following model:

$$\log u = \gamma_1 + \gamma_2 \log y + \tfrac{1}{2}\gamma_3 (\log\ y)^2 + \varepsilon,\ \varepsilon \sim N(0, \sigma_u^2). \quad (125)$$

Our prior on β and γ is flat and the priors on σ_v and σ_u are the same as before. Marginal posteriors of the inefficiency effect parameters (γ_j) are reported in Fig. 1. From the marginal posteriors, it turns out that these parameters are reasonably from zero. The sample distribution of efficiency is presented in Fig. 2. Electric utilities operate at relatively high levels of

[5]See Table F4.4 in http://pages.stern.nyu.edu/~wgreene/Text/Edition7/tablelist8new.htm which contains data sets for W. Greene, Econometric Analysis, 8th edition, Pearson, 2018.

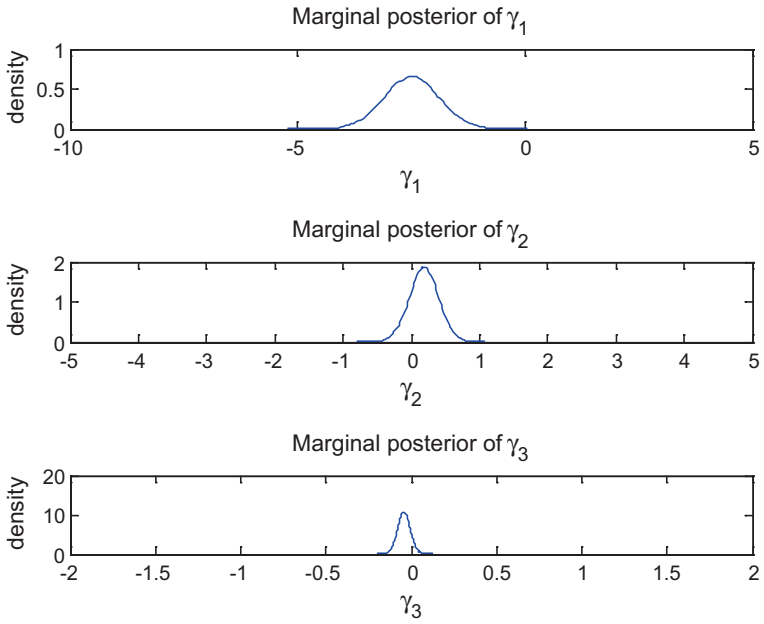

Fig. 1 Marginal posterior densities of inefficiency effect parameters

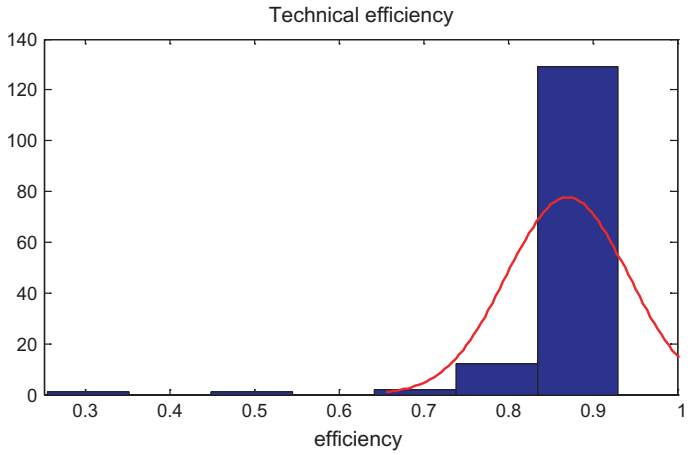

Fig. 2 Sample distribution of efficiency estimates

efficiency; average efficiency is close to 0.9 and ranges from about 0.65 to 1. The interesting feature is that there is considerable probability mass in the neighborhood of one.

Acknowledgements The author is indebted to an anonymous reviewer for comments on an earlier version of this chapter. Dedicated to John Geweke, Bill Greene, and Subal Kumbhakar for all that they taught me throughout the years.

References

Assaf, A.G., H. Oh, and M.G. Tsionas. 2017. Bayesian approach for the measurement of tourism performance: A case of stochastic frontier models. *Journal of Travel Research* 56 (2): 172–186.

Atkinson, S.C., and M.G. Tsionas. 2016. Directional distance functions: Optimal endogenous directions. *Journal of Econometrics* 190: 301–314.

Chib, S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90: 1313–1321.

Cornwell, C., P. Schmidt, and R. Sickles. 1990. Production Frontiers with cross sectional and time series variation in efficiency levels. *Journal of Econometrics* 46: 185–200.

DiCiccio, T.J., R.E. Kass, A. Raftery, and L. Wasserman. 1997. Computing Bayes factors using simulation and asymptotic approximations. *Journal of the American Statistical Association* 92: 903–915.

Fernandez, C., G. Koop, and M.F.J. Steel. 2002. Multiple-output production with undesirable outputs. *Journal of the American Statistical Association* 97: 432–442.

Fernandez, C., J. Osiewalski, and M.F.J. Steel. 1997. On the use of panel data in stochastic frontier models with improper priors. *Journal of Econometrics* 79: 169–193.

Fried, H.O., C.A.K. Lovell, and S.S. Schmidt. 1993. *The measurement of productive efficiency and productivity growth*. Oxford: Oxford University Press.

Gandhi, A., S. Navarro, and D. Rivers. 2013. On the identification of production functions: How heterogeneous is productivity? Working Paper.

Geweke, J. 1989. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57: 1317–1339.

Geweke, J. 1991. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian statistics 4*, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford: Oxford Press.

Geweke, J. 1999. Using simulation methods for Bayesian econometric models: Inference, development, and communication. *Econometric Reviews* 18 (1): 1–73.

Greene, W. 1990. A gamma-distributed stochastic frontier model. *Journal of Econometrics* 46: 141–163.

Greene, W. 2005. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics* 126 (2): 269–303.

Griffin, J. 2011. Bayesian clustering of distributions in stochastic frontier analysis. *Journal of Productivity Analysis* 36: 275–283.

Griffin, J.E., and M.F.J. Steel. 2004. Semiparametric Bayesian inference for stochastic frontier models. *Journal of Econometrics* 123: 121–152.

Koop, G., J. Osiewalski, and M.F.J. Steel. 1997. Bayesian efficiency analysis through individual effects. *Journal of Econometrics* 76: 7–105.

Koop, G., M.F.J. Steel, and J. Osiewalski. 1995. Posterior analysis of stochastic frontiers models using Gibbs sampling. *Computational Statistics* 10: 353–373.

Kumbhakar, S.C. 1997. Modeling allocative inefficiency in a translog cost function and cost share equations: An exact relationship. *Journal of Econometrics* 76: 351–356.

Kumbhakar, S.C., and E.G. Tsionas. 2005a. Measuring technical and allocative inefficiency in the translog cost system: A Bayesian approach. *Journal of Econometrics* 126 (2): 355–384.

Kumbhakar, S.C., and E.G. Tsionas. 2005b. The Joint Measurement of Technical and Allocative Inefficiencies: An application of Bayesian inference in nonlinear random-effects models. *Journal of the American Statistical Association* 100: 736–747.

Kumbhakar, S.C., B.U. Park, L. Simar, and M.G. Tsionas. 2007. Nonparametric stochastic frontiers: A local maximum likelihood approach. *Journal of Econometrics* 137 (1): 1–27.

Kutlu, L. 2010. Battese-Coelli estimator with endogenous regressors. *Economics Letters* 109 (2): 79–81.

Levinsohn, J., and A. Petrin. 2003. Estimating production functions using inputs to control for unobservables. *Review of Economic Studies* 70 (2): 317–341.

Marschak, J., and W.H. Andrews. 1944. Random simultaneous equations and the theory of production. *Econometrica* 12 (3/4): 143–205.

Olley, G.S., and A. Pakes. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64 (6): 1263–1297.

Simar, L., and W.P. Wilson. 1998. Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science* 44 (1): 49–61.

Simar, L., and W.P. Wilson. 2000. Statistical inference in nonparametric frontier models: The state of the art. *Journal of Productivity Analysis* 13: 49–78.

Simar, L., and W.P. Wilson. 2004. Performance of the bootstrap for DEA estimators and iterating the principle. In *Handbook on Data Envelopment Analysis*, ed. W.W. Cooper, M.L. Seiford, and J. Zhu, 265–298. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Tanner, M.A., and W.H. Wong. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82: 528–540.

Tran, K., and M.G. Tsionas. 2013. GMM estimation of stochastic frontier model with endogenous regressors. *Economics Letters* 108 (1): 233–236.

Tran, K., and M.G. Tsionas. 2015. One the estimation of zero-inefficiency stochastic frontier models with endogenous regressors. *Economics Letters* 147: 19–22.

Tran, K., and M.G. Tsionas. 2016. Endogeneity in stochastic frontier models: Copula approach without external instruments. *Economics Letters* 133: 85–88.

Tsionas, E.G. 2000. Full likelihood inference in normal-gamma stochastic frontier models. *Journal of Productivity Analysis* 13 (3): 183–205.

Tsionas, E.G. 2006. Inference in dynamic stochastic frontier models. *Journal of Applied Econometrics* 21 (5): 669–676.

Tsionas, E.G. 2012. Maximum likelihood estimation of stochastic frontier models by the Fourier transform. *Journal of Econometrics* 170: 234–248.

Tsionas, M.G. 2017. "When, where and how" of efficiency estimation: Improved procedures for stochastic frontier models. *Journal of the American Statistical Association* 112 (519): 948–965.

Tsionas, E.G., and S.C. Kumbhakar. 2012. Firm heterogeneity, persistent and transient technical inefficiency: A generalized true random-effects model. *Journal of Applied Econometrics* 29: 110–132.

van den Broeck, J., G. Koop, J. Osiewalski, and M.F.J. Steel. 1994. Stochastic frontier models: A Bayesian perspective. *Journal of Econometrics* 61: 273–303.

Common Methodological Choices in Nonparametric and Parametric Analyses of Firms' Performance

Luis Orea and José Luis Zofío

Once you choose, it is path-dependent

1 Introduction

This chapter is concerned with the choice of methods related to the theory and practice of productive and economic efficiency analysis. A methodological text such as this handbook, intended as a guide to frontier analysis, always benefits from a discussion on the common methodological, theoretical and empirical choices that scholars face when undertaking research in the field. We focus on the different forks that practitioners encounter. These range from the selection of the appropriate economic model to the use of the empirical techniques best suited to achieve results with confidence.

As departing point, and since this handbook is mainly concerned with the economic side of management practice, exceeding the engineering issues related to production processes, we hereafter consider the firm as the

L. Orea (✉)
Oviedo Efficiency Group, Department of Economics,
School of Economics and Business, University of Oviedo, Oviedo, Spain
e-mail: lorea@uniovi.es

J. L. Zofío
Department of Economics, Universidad Autónoma de Madrid,
Madrid, Spain
e-mail: jose.zofio@uam.es

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_12

relevant decision unit, operating within markets. When market decisions are involved, prices are key in the analysis, and we are in the realm of business and economics. This justifies the theoretical focus of the chapter on the concept of overall economic efficiency (e.g. profit or profitability efficiency), which starts in the following Sect. 2 by summarizing the main results of duality theory. Subsequently, once the theoretical framework has been decided, the next question relates to the choice of the most suitable methods to characterize the production technology, economic behaviour and, finally, measure firms' performance. Section 3 outlines the most popular empirical methods available to undertake efficiency analyses, namely nonparametric data envelopment analysis (DEA) and parametric stochastic frontier analysis (SFA). We discuss key issues related to imposing alternative technological assumptions and properties that, depending on the economic objective of the firm, require alternative specifications, e.g. returns to scale as well as quantity and price conditions of homogeneity.

Section 4 deals with the management of dimensionality difficulties in empirical research. The increased availability of large data sets including many variables compromises the reliability of results and reduces the discriminatory power of the efficiency analysis when the number of observations is limited. We discuss several strategies aimed at reducing the dimensionality of the analysis, either by relying on dimension reduction techniques that aggregate the original variables into a smaller set of composites, or by selecting those that better characterize production and economic processes. Another critical issue in the literature, discussed in Sect. 5, is the need to control for environmental or contextual z-variables that do not fall within managerial discretion. Nondiscretionary variables have both been included as frontier regressors or as determinants of firms' inefficiency and we discuss the implications that choosing each option has for researchers, managers and policy makers.

The fact that some variables may be endogenous or exhibit a large correlation with firms' inefficiency is gaining increasing attention in the literature. Section 6 presents a series of recent models addressing this issue in the DEA and SFA approaches. We summarize the main features of these methods and identify their relative advantages and disadvantages. In this section, we also discuss the endogenous nature of distance function when assessing firms' efficiency. The specific choice of orientation is also considered, as well as its ultimate relation to the concept of Pareto efficiency, i.e. utility maximization, which for the competitive firm results in the initial profit or profitability maximization objectives. For both DEA and SFA approaches, we also

present data-driven models that allow identifying individual directions based on local proximity (comparability) criteria. Finally, Sect. 7 summarizes the guiding principles of the chapter and draws the main conclusions.[1]

2 Theoretical Background: Firms' Objective and Decision Variables

2.1 Distance Functions, Economic Behaviour, Duality and Efficiency

We first introduce several technical (primal) representations of firms' technology. Next, we outline duality theory that relates these primal representations with a supporting (dual) economic function, capable of providing a consistent framework for the decomposition of economic efficiency into technical and allocative criteria. In economic efficiency analysis, the technical dimension of the problem is approached via the optimizing behaviour of the firm, either as cost minimizer, or as revenue, profit or profitability maximizer. More details can be found in Färe and Primont (1995), who summarize duality theory from an input (cost) and output (revenue) perspective, Chambers et al. (1996, 1998) for the profit function and the directional distance function (DDF), and Zofío and Prieto (2006), focusing on the relationship between the profitability (return-to-dollar) function and the generalized distance function (GDF).[2] A generalization of these relationships based on the *loss function* is introduced by Aparicio et al. (2016). Depending on the features of the study (e.g. market structure, firms' economic goal, data availability, etc.), researchers may choose any of these dimensions for their economic efficiency analysis.

[1]For an extended and augmented version of this chapter, the reader is referred to Orea and Zofío (2017).

[2]The DDF by Chambers et al. (1996) corresponds to the concept of shortage function introduced by Luenberger (1992, p. 242, Definition 4.1), which measures the distance of a production plan to the boundary of the production possibility set in the direction of a vector g. In other words, the shortage function measures the amount by which a specific plan falls short of reaching the frontier of the technology. Chambers et al. (1996) redefine the shortage function as efficiency measure, introducing the concept of DDF.

2.2 The Multi-output, Multi-input Production Technology: Distance Functions

Duality theory departs from the characterization of the technology set: $T = \{(x,y) : x \in \mathbb{R}^N_+, y \in \mathbb{R}^M_+,\ x \text{ can produce } y\}$ where x is a vector of input quantities, y is a vector of output quantities, and N and M are the number of inputs and outputs. The technology satisfies the axioms presented in Färe and Primont (1995): closedness, free disposability, no free-lunch and convexity. If the technology exhibits constant returns to scale (CRS), then the corresponding set is a cone, denoted by $T^{CRS} = \{(\psi x, \psi y) : (x,y) \in T,\ \psi > 0\}$.[3] This formula may also be applied to a technology T with variable returns to scale, extending it to the smallest cone containing T. This technique is useful to measure scale efficiency. For the single-output case: $M = 1$, the technology can be represented in what is termed as the primal approach by the production function $f : \mathbb{R}^N_+ \to \mathbb{R}_+$, defined by $f(x) = \max\{y : (x,y) \in T\}$, i.e. the maximum amount of output that can be obtained from any combination of inputs. The advantage of this interpretation is that it leaves room for technical inefficiency, since we can define a technology set parting from the production function by $T = \{(x,y) : f(x) \geq y, y \in \mathbb{R}_+\}$. Nevertheless, in the general (and real) multiple-output-multiple-input case, a suitable representation of the technology is given by the distance function introduced by Shephard (1970).

This representation can be made from alternative orientations. Here, we consider recently introduced and rather flexible characterizations of the technology corresponding to the *additive directional* and *multiplicative generalized* distance functions (DFs):

- The directional DF:

$$D_T(x,y;-g_x,g_y) = \max\{\beta : (x - \beta g_x, y + \beta g_y) \in T\}, \tag{1}$$

- The generalized DF:

$$D_G(x,y;\alpha) = \min\left\{\delta : (\delta^{1-\alpha}x, y/\delta^{\alpha}) \in T\right\}. \tag{2}$$

[3]In empirical studies approximating the technology through DEA, the global CRS characterization is assumed for analytical convenience because relevant definitions such as profitability efficiency and the Malmquist productivity index require this scale property, and therefore, their associated distance functions are defined with respect to that benchmark technology.

The DDF, $D_T\left(x, y; -g_x, g_y\right) \geq 0$, is a measure of the maximal translation of (x, y) in the direction defined by $\left(g_x, g_y\right) \in \mathbb{R}_+^N \times \mathbb{R}_+^M \backslash \{0_{N+M}\}$ that keeps the translated input–output combination inside the production possibility set. The GDF, $0 < D_G(x, y;\ \alpha) \leq 1$, rescales (x, y) according to the parameter $0 \leq \alpha \leq 1$, also keeping the projected input–output combination inside the production possibility set. The properties of these functions are presented in Chambers et al. (1996, 1998) and Chavas and Cox (1999). More importantly, one of their most relevant features is that they nest Shephard's input and output distance functions depending on the specific values of the directional vector $g \neq 0$ or directional parameter α. The input distance function is obtained by setting $g = \left(-g_x, g_y\right) = (-x, 0)$ and $\alpha = 0$, while the output distance function corresponds to $g = \left(-g_x, g_y\right) = (0, y)$ and $\alpha = 1$.[4] Additionally, the GDF is the only one which nests the hyperbolic distance function introduced by Färe et al. (1985) for $\alpha = 0.5$.[5] This implies that both approaches can generalize Shephard's input and output distance functions, and therefore, their inverse corresponds to Farrell's (1957) radial efficiency measures. However, we note in what follows that such a generalization for the case of the DDF does not extend to the notions of cost and revenue efficiency and its decomposition into technical and allocative efficiency, since the latter does not verify a dual relationship when characterizing the technology through the DDF as shown by Aparicio et al. (2017).

The choice of direction by the researcher, addressed in Sect. 6.3, represents an initial challenge. Settling for an input or output orientation restricts the production or economic analysis to one dimension (cost or revenue), while allowing for alternative directions requires justification, including those that assign different directions for each observation. Although the aforementioned distance functions rely on the same set of variables, and

[4]The *input* and *output* distance functions define respectively as $D_I(x, y) = \max\{\lambda : (x/\lambda, y) \in T\}$ and $D_O(x, y) = \min\{\theta : (x, y/\theta) \in T\}$. If the technology satisfies the customary axioms, the input distance function has the range $D_I(x, y) \geq 1$. It is homogeneous of degree one in inputs, non-decreasing in inputs and nonincreasing in outputs. In contrast, the output distance function has the range $0 < D_O(x, y) \leq 1$. It is homogeneous of degree one in outputs, nondecreasing in outputs and nonincreasing in inputs. Färe and Primont (1995, pp. 15, 22) show that weak disposability of inputs and outputs is necessary and sufficient for the input and output distance functions to completely characterize technology.

[5]The *hyperbolic* distance function inherits its name from the hyperbolic path that it follows towards the production frontier. The range of the hyperbolic distance function is $0 < D_H(x, y) \leq 1$. It satisfies the following properties: it is almost homogeneous of degrees k^1, k^2 and k^3: $D_H(\lambda^{k_1} x, \lambda^{k_2} y; \alpha) = \lambda^{k_3} D_H(x, y; \alpha)$, for all $\lambda > 0$, $k = (-1, 1, 1)$ (Aczél 1966, Chs. 5 and 7; Cuesta and Zofío 2005), nondecreasing in outputs and nonincreasing in inputs.

sometimes share the same parametric representation, the question that naturally arises is which formulation should be used in empirical applications. An economic criterion is needed to choose among these alternatives. Debreu's (1951) "coefficient of resource utilization" suggests Pareto efficiency, based on utility maximization by economic agents.[6] At the firm level, this implies profit maximization at competitive market prices, which in turn entails cost minimization. Additionally, under CRS, this is equivalent to profitability maximization.

2.3 Optimizing Economic Behaviour

The following economic objectives allow us to discuss the duality framework for an overall economic efficiency analysis. Based on the previous primal representations of the technology (1) and (2), and considering the vectors of input and output shadow prices, $w \in \mathbb{R}^N_+$ and $p \in \mathbb{R}^M_+$, the following economic functions can be defined:

- The profit function:

$$\pi(w,p) = \max_{x,y} \{py - wx : (x,y) \in T\}, \tag{3}$$

- The profitability function:

$$\rho(w,p) = \max_{x,y} \left\{ py/wx : (x,y) \in T^{CRS} \right\}. \tag{4}$$

The profit function determines the maximal feasible profit defined as revenue minus cost, and assuming the necessary derivative properties—including continuity and differentiability, Hotelling's lemma yields the input demand and output supply functions. Alternatively, the profitability or return-to-dollar (RD) function represents the maximum attainable revenue to cost ratio.[7]

[6]Debreu's (1951) "coefficient of resource utilization" is the corner stone upon which Aparicio et al. (2016) introduce the concept of *loss distance function*, identifying the minimum conditions necessary to derive a dual relationship with a supporting economic function. They obtain specific normalizing sets of the loss function that correspond to the most usual distance functions.

[7]The counterpart to the input distance function corresponds to the cost function, defined as $C(y,w) = \min_x \{wx : x \in L(y)\}$, where $L(y) = \{x : (x,y) \in T\}$ is the input requirement set. It represents the minimum cost of producing a given amount of outputs, yielding the input demand functions by applying Shephard's lemma. Correspondingly, the revenue function $R(x,p) = \max_y \{py : y \in P(x)\}$,

For the optima (min or max) to exist, conditions must be fulfilled. In the case of the profit function, nonincreasing returns to scale are required (with profit equalling 0 or $+\infty$ under CRS). For the profitability function, Zofío and Prieto (2006) prove that maximum profitability is attained in *loci* where the production technology exhibits local CRS—i.e. processes exhibiting a technically optimal scale, Balk (1998, p. 19), and constituting a most productive scale size (MPSS) in Banker and Thrall's (1992) terminology. A suitable GDF intended to measure overall economic efficiency, including scale efficiency, is relative to a production possibility set with CRS, using as a benchmark the virtual cone technology, T^{CRS}.[8]

2.4 Duality and Overall Economic Efficiency: Technical and Allocative Efficiency

Several authors, including Chambers et al. (1998) for (3) and Zofío and Prieto (2006) for (4), prove the duality between the aforementioned distance functions and their associated economic functions. In particular, interpreting the distance functions as measures of technical efficiency, it is possible to define the following inequalities[9]:

- Profit:

$$T = \{(x,y) : py - wx + D_T\left(x,y;-g_x,g_y\right)\left(pg_y + wg_x\right) \leq \pi(w,p) \text{ for all } p,w > 0\} \tag{5}$$

where $P(x) = \{y : (x,y) \in T\}$ is the output production possibility set, represents the maximum possible revenue of using a given amount of inputs, yielding the output supply functions.

[8]The technology may be characterized by variable returns to scale as in (2), allowing for scale (in)efficiency $D_G^{CRS}(x,y;\alpha) = D_G(x,y;\alpha)SE_G$, with $SE_G = D_G^{CRS}(x,y;\alpha)/D_G(x,y;\alpha)$, but the final supporting technological benchmark is characterized by CRS.

[9]Here, we take into account that $T = \{(x,y) : D_G(x,y;\alpha) \leq 1\}$ and $T = \left\{(x,y) : D_T\left(x,y,-g_x,g_y\right) \geq 0\right\}$. For the case of the profit and DDFs, the additive overall efficiency measure is normalized by $pg_y + wg_x = 1$, ensuring that it is independent of the measurement units as its multiplicative counterparts—see Nerlove (1965). These dual relations are economic particularizations of Minkowski's (1911) theorem: every closed convex set can be characterized as the intersection of its supporting halfspaces. In fact, the cost, revenue, profit and profitability functions are known as the support functions characterizing the technology for alternative shadow prices—e.g. for the particular case of the cost function, see Chambers (1988, p. 83).

- Profitability:

$$T^{CRS} = \{(x,y) : p\left(y\, D_G^{CRS}(x,y;\alpha)^{-\alpha}\right)/w\left(x\, D_G^{CRS}(x,y;\alpha)^{1-\alpha}\right) \leq \rho(w,p) \text{ for all } p,w > 0\}. \quad (6)$$

Closing the inequalities by adding a residual variable capturing allocative inefficiency, allows establishing the following decompositions of overall economic efficiency:

- Overall profit (Nerlovian) inefficiency:

$$\frac{\pi(w,p) - (py - wx)}{pg_y + wg_x} = D_T\left(x,y;-g_x,g_y\right) + AI_T, \quad (7)$$

- Overall profitability (RD) efficiency:

$$\frac{py/wx}{\rho(w,p)} = D_G^{CRS}(x,y;\alpha)AE_G^{CRS}. \quad (8)$$

The above relationships constitute the core of the empirical research on productive and economic efficiency when market prices are brought into the analysis.[10] We can now define overall economic efficiency as the ability of firms to achieve their economic goal, either maximum profit or profitability, which in turn requires that they are technically efficient by using the available technology at its best, as well as allocative efficient by demanding and supplying the right mix of inputs and outputs. Recalling the notion of optimality that opens this section, the underlying concept is again that of Pareto efficiency. Accordingly, and following Koopmans (1951), it is possible to state the following definition of technical efficiency: a firm is technically efficient if an increase in any output requires a reduction in at least one other output or an increase in at least one input and if a reduction in any input requires an increase in at least one other input or a reduction in at least one output. In formal terms, this definition implies that the firm belongs to the strongly efficient subset of the technology:

[10]The overall cost and revenue efficiencies correspond to $C(y,w)/wx = (1/D_I(x,y)) \cdot AE_I$ and $py/R(x,p) = D_O(x,y) \cdot AE_O$, respectively.

$Eff(T) = \{(x,y) \in T : (u,-v) \leq (x,-y), (u,v) \neq (x,y) \Rightarrow (u,v) \notin T\}$.[11] The distance functions (7) and (8) represent suitable measures of technical efficiency, albeit referred to weaker notions of efficiency characterized by their corresponding subsets—e.g. for the particular case of the DDF, see Aparicio et al. (2016, p. 76). Values of $D_T\left(x,y;-g_x,g_y\right) = 0$ and $D_G^{CRS}(x,y;\alpha) = 1$ signal that, given the technology, simultaneous changes in inputs and outputs so as to improve productive efficiency are infeasible. A definition of allocative efficiency can also be provided in terms of market prices: A firm is allocative efficient if it demands the optimal amounts of inputs and supplies the optimal amounts of outputs that maximize either profit or profitability at the existing prices. In the former case, the value of allocative efficiency is a residual defined as the difference between profit efficiency and the DDF, while in the latter case its value corresponds to the ratio of profitability efficiency to the GDF. For $AI_T = 0$ and $AE_G^{CRS} = 1$, the firm is allocative efficient.

Several remarks are relevant for applied research. First note that for the overall profitability decomposition, the CRS benchmark characterizes the GDF. Second, a less restrictive property, homotheticity, is also required for a meaningful decomposition of overall economic efficiency, where the distance functions can be rightly interpreted as measures of technical efficiency. Within a nonparametric setting, Aparicio et al. (2015) and within a parametric setting Aparicio and Zofío (2017) show that, for nonhomothetic technologies, the radial contractions (expansions) of the input (output) vectors resulting in efficiency gains do not maintain allocative (in)efficiency constant along the firm's projection to the production frontier (isoquants). This implies that they cannot be solely interpreted as technical efficiency reductions. From the perspective of, for example, the cost and revenue efficiency decompositions, this result invalidates the residual nature of allocative efficiency and requires the use of a distance function with a directional vector capable of keeping allocative efficiency constant along the projections.[12] Third, while the additive DDF nests the input and output radial distance functions, such generalization does not extend to the notion of cost or revenue efficiency and its decomposition into technical and allocative terms. For these particular directions, allocative efficiency cannot be obtained as an

[11]The (strongly) efficient set consists of all firms that are not dominated, requiring monotonic preferences to characterize efficiency (ten Raa 2008, p. 194, Lemma).

[12]This in turn implies that the radial framework or choosing as a directional vector the observed amounts of inputs and outputs in the case of the DDF is no longer valid.

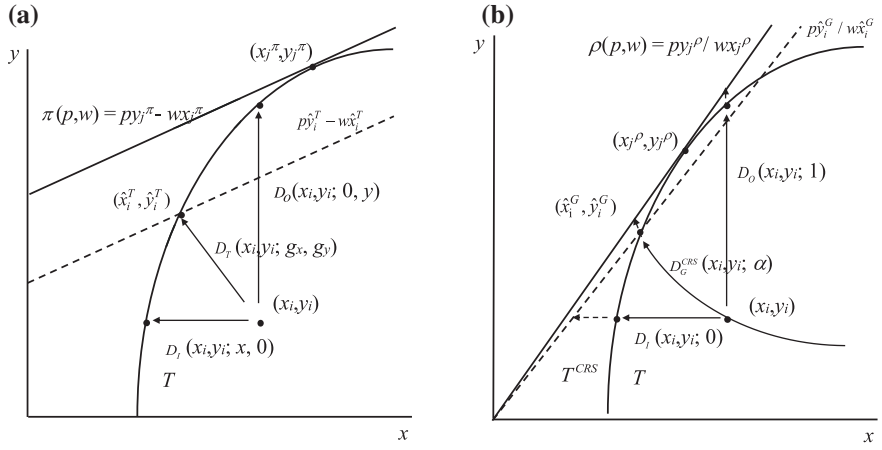

Fig. 1 Distance functions and their economic duals: Profit (**a**) and profitability (**b**)

independent residual from the above inequalities as shown by Aparicio et al. (2017).[13]

The alternative distance function representations of production technology (technical efficiency measures), dual economic functions and the residual nature of allocative efficiency are presented in Fig. 1. We comment overall economic efficiency for the directional and GDFs and their dual profit and profitability functions. In the left panel, (a) the directional function (1) measuring the distance from the single-input-single-output unit (x_i, y_i) to the frontier is represented by $D_T\left(x_i, y_i;\, g_x,\, g_y\right)$, measuring technical inefficiency and, equivalently—thanks to the duality relationship (5)—its associated profit loss in monetary units if the normalizing constraint is set to $\left(pg_y + wg_x\right) = 1$. This projects the unit to $\left(\hat{x}_i^T, \hat{y}_i^T\right)$, whose profit is $p\hat{y}_i^T - w\hat{x}_i^T$. Therefore, and thanks to (7), the difference between maximum profit—attained at $\left(x_i^{\pi}, y_i^{\pi}\right)$—and observed profit at the efficient projection corresponds to allocative inefficiency: $\pi(p,w) - \left(p\hat{y}_i^T - w\hat{x}_i^T\right)$. In the same panel, (a) the input and output distance functions are also presented as particular cases of the directional formulation for $\left(-g_x, g_y\right) = (-x,\, 0)$ and $\left(-g_x, g_y\right) = (0,\, y)$, but whose interpretation in terms of overall cost or

[13]Regarding denominations, we note that a firm is overall profit efficient when its technical and allocative terms are zero rather than one. This implies that the larger the numerical value of the DDF the more inefficient is the firm, thus the technical and allocative (*in*)efficiency notation: *TI* and *AI*, with $TI = D_T\left(x, y;\, -g_x, g_y\right)$. Other authors, e.g. Balk (1998), favour a consistent characterization of efficiency throughout, so the larger the value the greater the firm's efficiency. This is achieved by defining $TE = -D_T\left(x, y;\, -g_x, g_y\right)$.

revenue efficiency is inconsistent. The right panel (b) presents an equivalent analysis in terms of the GDF (2) projecting the evaluated unit to $(\hat{x}_i^G, \hat{y}_i^G)$ through $D_G^{CRS}(x, y;\ \alpha)$. In the single-input-single-output case, its technical inefficiency interpretation is the amount by which observed average productivity y_i/x_i can be increased to attain, $\hat{y}_i^G/\hat{x}_i^G$ at the reference frontier. Now, thanks to the duality relationship (6), the difference can be interpreted in terms of profitability differentials given the input and output market prices. Finally, following (8), it is possible to determine allocative efficiency as the ratio between projected profitability $p\hat{y}_i^G/w\hat{x}_i^G$ to maximum profitability—attained at (x_i^ρ, y_i^ρ): i.e. $\left(p\hat{y}_i^G/w\hat{x}_i^G\right)/\left(py_i^\rho/wx_i^\rho\right)$. Again, in the same panel (b) the input and output distance functions are presented for $\alpha = 0$ and $\alpha = 1$, respectively.

3 Standard Approaches to Measure Firms' Economic Efficiency

Once the theoretical framework has been introduced, the next step is the consideration of the empirical methods that allow the measurement and decomposition of firms' economic efficiency relying on duality theory. This requires approximating the technology by using either nonparametric mathematical programming, parametric econometric techniques (regression analysis) or engineering (bottom-up) models. In this section, we only describe the main features of the two most popular approaches, DEA and SFA. Throughout the section, we limit our discussion to simple specifications of both parametric and nonparametric models.[14]

3.1 Data Envelopment Analysis

DEA approximates the production technology from observed data by relying on the principle of minimum extrapolation, which yields the smallest subset of the input–output space $\mathbb{R}_+^N \times \mathbb{R}_+^M$ containing all observations, and satisfying certain technological assumptions. Technology consists of piecewise linear combinations of the observed $i = 1, \ldots, I$ firms constituting a

[14]A comprehensive exposition is presented in earlier chapters devoted to the deterministic and stochastic benchmarking methodologies by Subash Ray, and William H. Greene and Phill Wheat, respectively.

multidimensional production frontier.[15] The most common DEA piecewise linear approximation of the technology T is given by

$$T = \left\{ (x,y) \in \mathbb{R}_+^N \times \mathbb{R}_+^M : \sum_{i=1}^{I} \lambda_i x_{in} \leq x_n, n = 1,\ldots,N; \right.$$
$$\left. \sum_{i=1}^{I} \lambda_i y_{im} \geq y_m, m = 1,\ldots,M; \sum_{i=1}^{I} \lambda_i = 1, \lambda \in \mathbb{R}_+^I, i = 1,\ldots,I \right\} \tag{9}$$

where λ is an intensity vector whose values determine the linear combinations of *facets* which define the production frontier, and whose associated restrictions allow considering alternative returns to scale; e.g., $\sum_i \lambda = 1$ implies variable returns to scale. Among the technological properties incorporated into the above DEA model, we highlight convexity, strong disposability and variable returns to scale.

Regarding *convexity*, while there are alternative DEA models dropping this assumption like the *Free Disposal Hull* (FDH) or *Free Replicability Hull* (FRH), these are inconsistent with duality theory (i.e. Minkowski's theorem), since convexity is key when recovering the technology from the supporting economic functions. As for *free (or strong) disposability*, implying that it is feasible to discard unnecessary inputs and unwanted outputs without incurring in technological opportunity costs is a rather weak assumption that, nevertheless, has its drawbacks. Most importantly, when measuring technical efficiency through radial distance functions, their values reflect whether the firm belongs to the so-called isoquant subsets—again a notion of efficiency weaker than the previously considered by Koopmans (1951) and that leaves room for nonradial efficiency improvements associated with strong disposability (i.e. slacks). Finally, alternative *returns to scale* can be postulated in (9) through the intensity variables λ. The variable returns to scale assumption can be dropped in favour of CRS by removing $\sum_{i=1}^{I} \lambda_i = 1$, while nonincreasing and nondecreasing returns to scale correspond to $\sum_{i=1}^{I} \lambda_i \leq 1$ and $\sum_{i=1}^{I} \lambda_i \geq 1$, respectively.

After characterizing the technology, the distance functions (1) and (2) can be calculated by solving their corresponding mathematical programmes. Again, we present only the directional and GDFs, as their input, output and hyperbolic distance functions are particular cases. Taking as guiding framework the decomposition of economic efficiency and the scale properties of

[15]See Cooper et al. (2007) and Färe et al. (1994) for an introduction to the Activity Analysis DEA within a production theory context.

the technology associated with either profit or profitability maximization, respectively, corresponding to nonincreasing and CRS, the following programmes allow evaluating the technical efficiency of firm $(x_{i'}, y_{i'})$:

Directional distance function, DDF

$$\begin{aligned} & D_T\left(x_{i'}, y_{i'}; -g_x, g_y\right) = \\ & = \max_{\beta,\lambda_i} \left\{\beta : \left(x_{i'} - \beta g_x, y_{i'} + \beta g_y\right) \in T\right\} \\ \text{s.t.} \quad & \sum_{i=1}^{\mathrm{I}} \lambda_i x_{in} \leq x_{i'n} - \beta g_{x_n},\ n = 1, \ldots, N, \\ & \sum_{i=1}^{\mathrm{I}} \lambda_i y_{im} \geq y_{i'm} + \beta g_{y_m},\ m = 1, \ldots, M, \\ & \sum_{i=1}^{\mathrm{I}} \lambda_i \leq 1, \quad \lambda \in \mathbb{R}_+^I. \end{aligned} \tag{10}$$

Generalized distance function, GDF

$$\begin{aligned} & D_G^{CRS}(x_{i'}, y_{i'};\ \alpha) = \\ & = \min_{\delta,\lambda_i} \left\{\delta : \left(x_{i'}\delta^{1-\alpha}, y_{i'}/\delta^{\alpha}\right) \in T^{CRS}\right\} \\ \text{s.t.} \quad & \sum_{i=1}^{\mathrm{I}} \lambda_i x_{in} \leq \delta^{1-\alpha}\, x_{i'n}, \quad n = 1, \ldots, N, \\ & \sum_{i=1}^{\mathrm{I}} \lambda_i y_{im} \geq y_{i'm}/\delta^{\alpha}, \quad m = 1, \ldots, M, \\ & \lambda \in \mathbb{R}_+^I. \end{aligned} \tag{11}$$

Note that these programmes incorporate the DEA production possibility set (9) as technological restrictions.

Besides the values of the distance functions representing the technical efficiency measures, relevant information can be obtained from both the above "envelopment" formulations of the technology and their "multiplier" duals—see (12) and (13) below. As identified in (10) and (11), a firm can indeed compare its technical and economic performance to that of its peers, so as to improve its own efficiency and productivity; those firms with optimal $\lambda_i^* > 0$ conform the reference frontier, and the value corresponds to

the relevance (weight) of the benchmark firm in the linear combination. Regarding technical efficiency, the number of possible peer firms is equal to the number of inputs plus the number of outputs except in the CRS case, where there can generally be one less reference peer. This follows from Carathéodory's theorem stating that if a problem with n restrictions has an optimal solution, then there exists a solution in which at most n variables are positive—i.e. known as the basic feasible solution.

The duals corresponding to the above envelopment formulations of the directional and GDFs are the following[16]:

Directional distance function, DDF

$$\begin{aligned} &\mathrm{D}_T\left(x_{i'}, y_{i'}, -g_x, g_y\right) = \\ &\quad \min_{\mu,\nu,\omega} \nu x_{i'} - \mu y_{i'} + \omega \\ &\text{s.t.} \quad \nu x_i - \mu y_i + \omega \geq 0, \quad i = 1, \ldots, I \\ &\qquad \nu g_x + \mu g_y = 1, \\ &\qquad \nu \geq 0,\ \mu \geq 0,\ \omega \geq 0. \end{aligned} \tag{12}$$

Generalized distance function, GDF

$$\begin{aligned} &\mathrm{D}_G^{CRS}(x_{i'}, y_{i'}; \alpha) = \max_{\mu,\nu} \mu y_{i'} \\ &\text{s.t.} \quad -\nu\, x_i + \mu y_i \leq 0, \quad i = 1, \ldots, I \\ &\qquad \mu x_{i'} = 1, \\ &\qquad \nu \geq 0,\ \mu \geq 0. \end{aligned} \tag{13}$$

From these programmes, technological relationships between inputs and outputs can be discerned, in the form of the optimal weights or multipliers (ν, μ), defining the supporting (reference) hyperplanes against which technical efficiency is measured. In this case, the firm is efficient if it belongs to one of the supporting hyperplanes (forming the facets of the envelopment surface) for which all firms lie on or beneath it.

The choice of the primal "envelopment" formulations (10) and (11) or their "multiplier" duals (12) and (13) depends on the analytical objective of researchers and the specific characteristics of the study. Nevertheless, the simplex method for solving the envelopment form also produces the optimal values of the dual variables, and all existing optimization software provides

[16]The dual for the GDF envelopment formulation (11) can be determined because it corresponds to a CRS characterization of the production technology, rendering it equivalent, for instance, to the radially oriented output distance function for $\alpha = 1$—since the value of $D_G^{CRS}(x, y; \alpha)$ is independent of α.

both sets of results readily, so there is not any computational burden on a particular choice of model.[17] For peer evaluation and determination of the nature of returns to scale, the envelopment formulations are adequate, while the duals are required if one wants to set weight restrictions rather than to adhere to the "most favourable weights" that DEA yields by default (Thompson et al. 1986; Podinovsky 2015). Also, as optimal weights are not unique, one can define secondary goals in comparative analyses that, using cross-efficiency methods, also help to rank observations that are efficient in the standard (first stage) DEA (Sexton et al. 1986; Cook and Zhu 2015).

Once the distance functions measuring technical efficiency have been calculated, it is possible to determine overall profit and profitability efficiency by resorting to DEA, subject to the same technology. These programmes incorporate the restrictions characterizing returns to scale and jointly determine the maximum profit or profitability, depending on the choice of economic behaviour:

Profit efficiency

$$
\begin{aligned}
&\frac{\pi(w,p)-(py_{i'}-wx_{i'})}{pg_y+wg_x}=\max\nolimits_{\beta,\lambda_i,x,y}\varphi\\
\text{s.t.}\quad &\sum_{i=1}^{\mathrm{I}}\lambda_i x_{in}=x_n,\ n=1,\ldots,N,\\
&\sum_{i=1}^{\mathrm{I}}\lambda_i y_{im}=y_m,\ m=1,\ldots,M,\\
&\sum_{i=i}^{\mathrm{I}}\lambda_i\frac{py_i-wx_i}{pg_y+wg_x}=\frac{py_{i'}-wx_{i'}}{pg_y+wg_x}+\varphi,\\
&\sum_{i=1}^{\mathrm{I}}\lambda_i\le 1,\quad \lambda\in\mathbb{R}_+^I.
\end{aligned}
\tag{14}
$$

[17]Nevertheless, the computational effort of solving the envelopment problems grows in proportion to powers of the number of DMUs, *I*. As the number of DMUs is considerably larger than the number of inputs and outputs ($N+M$), it takes longer and requires more memory to solve the envelopment problems. We contend that except for simulation analyses and the use of recursive statistical methods such as bootstrapping, nowadays processing power allows calculation of either method without computational burdens.

Profitability efficiency

$$\begin{aligned}
&\frac{py_{i'}/wx_{i'}}{\rho(w,p)} = \min_{\phi,\lambda_i,x,y} \phi \\
&\text{s.t.} \quad \sum_{i=1}^{\mathrm{I}} \lambda_i x_{in} = x_n, n = 1,\ldots,N, \\
&\qquad \sum_{i=1}^{\mathrm{I}} \lambda_i y_{im} = y_m, m = 1,\ldots,M, \\
&\qquad \sum_{i=i}^{\mathrm{I}} \lambda_i \frac{wx_i}{py_i} = \phi \frac{wx_{i'}}{py_{i'}}, \\
&\qquad \lambda \in \mathbb{R}_+^I.
\end{aligned} \tag{15}$$

The decomposition of overall economic efficiency can then be completed by calculating allocative efficiency as a residual, i.e. solving for AI_T and AE_G^{CRS} in Eqs. (7) and (8). That is, using the optimal solutions obtained for the directional and GDFs, as well as the calculated profit and profitability efficiencies:

- Allocative (in)efficiency:

$$AI_T \equiv \frac{\pi(w,p) - (py - wx)}{pg_y + wg_x} - D_T(x,y;-g_x,g_y) = \varphi^* - \beta^*, \tag{16}$$

- Allocative efficiency:

$$AE_G^{CRS} \equiv \frac{py/wx}{\rho(w,p)} / D_G^{CRS}(x,y;\alpha) = \delta^*/\phi^*, \tag{17}$$

3.2 Stochastic Frontier Approach

In this section, we outline the main features of the standard econometric approach to measuring firms' inefficiency. For a comprehensive survey of this literature, see Kumbhakar and Lovell (2000), Fried et al. (2008), and Parmeter and Kumbhakar (2014). For notational ease, we have developed this section for cross-sectional data, except when a panel data framework is discussed.

In analogy to the DEA analysis, we confine our discussion to the estimation of technical efficiency using distance functions.[18] Thus, firm performance is evaluated by means of the following (general) distance function:

$$\ln D_i = \ln D(x_i, y_i, \beta) + v_i, \tag{18}$$

where the scalar y_i is the output of firm $i = 1, \ldots I$, x_i is a vector of inputs, $\ln D_i$ captures firm's technical efficiency, $\ln D(x_i, y_i, \beta)$ is a deterministic function measuring firm's technology, β is now a vector of technological parameters, and v_i is a two-sided noise term. In Eq. (18), we specify the distance function as being stochastic in order to capture random shocks that are not under the control of the firm. It can also be interpreted as a specification error term that appears when the researcher tries to model the firm's technology.

A relevant issue that should be addressed here is that while the dual representations of the technology in (3) and (4) are clearly identified in a parametric exercise by the different sets of both dependent and explanatory variables, this is not the case for the primal representations based on the distance functions in (1) and (2). At first sight, all of them are functions of the *same* vector of inputs and outputs. Thus, if we were able to estimate a function of inputs and outputs, say $\ln D(x_i, y_i, \beta)$, how do we ensure that we have estimated our preferred choice, say, an output distance function, and not an input distance function? Note also that, regardless the orientation of the distance function, the term measuring firms' inefficiency (i.e. $\ln D_i$) is not observed by the researcher, and thus, it cannot be used as a proper dependent variable to estimate (18). For identification purposes, we need to rely on the theoretical properties of distance functions. In particular, the key property for identification is the translation property for the DDF, the almost homogeneity condition for the GDF (and its particular case corresponding to the hyperbolic definition) and the linear homogeneity condition for the input and output distance functions. Identification works because each homogeneity condition involves different sets of variables. Although the underlying technology is the same, the coefficients of each distance function differ. In the case of output distance

[18]Although most early SFA applications used production functions, the distance function became as popular as the production functions since Coelli and Perelman (1996), who helped practitioners to estimate and interpret properly the distance functions. In addition, the distance functions can constitute the building blocks for the measurement of productivity change and its decomposition into its basic sources (see, e.g., Orea 2002). This decomposition can be helpful to guide policy if estimated with precision.

function, for example, linear homogeneity in outputs implies that the deterministic distance function $\ln D(x_i, y_i, \beta)$ can alternatively be rewritten as:

$$\ln D(x_i, y_i, \beta) = \ln D\left(x_i, y_i / y_{1i}, \beta\right) + \ln y_{1i}. \tag{19}$$

This specification immediately "produces" an observed dependent variable for the above model once (19) is inserted into (18). Indeed, rearranging terms, model (18) can be expressed as follows[19]:

$$\ln y_{1i} = -\ln D(x_i, y_i/y_{1i}, \beta) + v_i - u_i, \tag{20}$$

where $u_i = -\ln D_i \geq 0$ is a nonnegative random term measuring firms' inefficiency that can vary across firms. Note that this model can be immediately estimated econometrically once a particular functional form is chosen for $\ln D(x_i, y_i/y_{1i}, \beta)$, and u_i is properly modelled.

The input and generalized (hyperbolic) distance functions, as well as the DDF, deserve similar comments. While the standard radial distance functions are mainly estimated in the literature using the Translog specification, the Quadratic function is often used for the DDF. The reason is that the translation property can be easily imposed on this specification (see Färe et al. 2005). Both Translog and Quadratic functions are not only differentiable allowing for the estimation of shadow prices and output/input substitutability, but also provide a second-order approximation to a true, but unknown distance function (see Diewert 1971, pp. 481–507).[20] It is worth mentioning that inefficiency is measured in physical units if a Quadratic specification is used, and not in percentage terms as it happens if we use a traditional Translog specification. Both measures are correct, albeit they are simply using different approaches to measure the distance to the frontier. On the other hand, an interesting feature that is often overlooked is that the Quadratic specification is normally estimated once the variables are normalized with the sample means (see Färe et al. 2005, p. 480). As the normalized variables are unit free, in practice the estimated

[19]To obtain this equation, we have taken into account that the v_i and $-v_i$ have the same normal distribution.

[20]While the flexibility of the functional forms allows a more precise representation of the production technology and economic behaviour, it is prone to some drawbacks. For instance, Lau (1986) proved that flexibility is incompatible with global regularity if both concavity and monotonicity are imposed using standard econometric techniques. That is, imposing regularity conditions globally often comes at the cost of limiting the flexibility of the functional form. It should be pointed out, however, that it is possible to maintain *local* flexibility using Bayesian techniques. See Griffiths et al. (2000) and O'Donnell and Coelli (2005).

inefficiency scores can be interpreted as proportional changes in outputs and inputs, in the same fashion as in the standard radial distance functions.

Note also that the composed error term $\varepsilon_i = v_i - u_i$ in (20) comprises two independent parts, a noise term and an inefficiency term. They are likely to follow different distributions given their different nature. Indeed, it is conventionally assumed that v_i follows a symmetric distribution since random shocks and specification errors might take both positive and negative values. However, by construction, inefficient performance always produces a contraction in firms' output. For this reason, u_i is assumed to be nonnegative (and asymmetrically) distributed. This results in a composed error term ε_i that is asymmetrically distributed. As customary in the literature, it is also assumed throughout that both random terms are distributed independently of each other and of the input variable.

We now turn to explaining how to estimate the above frontier model. The estimation of the model involves both the parameters of the distance (production) function and the inefficiency. Even with very simple SFA models, the researcher has several estimation methods at hand and, in most applications, chooses only one. All have their own advantages and disadvantages. Equation (20) can first be estimated via *Maximum likelihood* (ML) once particular distributional assumptions on both random terms are made. ML is the most popular empirical strategy in the literature, but it relies on (perhaps strong) assumptions regarding the distribution of these terms and the exogenous nature of the regressors. A second method that we can choose is the *Method of Moments* (MM) approach, where all technological parameters of the production function are estimated using standard econometric techniques without making specific distributional assumptions on the error components. If, in addition, we are willing to compute the firms' efficiency scores with no distributional assumptions on the error components, we can follow the so-called *Distribution-Free Approach*, which includes the well-known *Corrected Ordinary Least Squares* (COLS) method for cross-sectional data and the CSS method (from Cornwell et al., 1990) in a panel data setting.

If Eq. (20) is estimated via ML, both technological parameters of the production function (β) and the structure of the two error components (i.e. the variance of v_i and u_i) are estimated simultaneously in a single stage. If the MM approach is chosen, an additional stage is involved. In the first stage, all technological parameters of the production function are estimated using standard econometric techniques, such as OLS or GMM. This stage is independent of distributional assumptions in respect of either error component. In the second stage of

the estimation procedure, distributional assumptions are invoked to obtain ML estimates of the parameter(s) describing the structure of the two error components, conditional on the first-stage estimated parameters.[21] Although the MM approach is much less used by practitioners than the traditional ML approach, the most comprehensive SFA versions of the MM estimator are becoming increasingly popular among researchers because it allows for instance dealing with endogenous variables (see Guan et al. 2009) or distinguishing between transient and permanent efficiency (Filippini and Greene 2016).

Whatever the approach we favour, we are forced to choose a distribution for v_i and u_i in order to estimate the parameters in Eq. (20) by ML. While the noise term is often assumed to be normally distributed with zero mean and constant standard deviation, several distributions have been proposed in the literature for the inefficiency term, viz. half-normal (Aigner et al. 1977), exponential (Meeusen and van den Broeck 1977) and gamma (Greene 1990). By far, the most popular distribution is the half-normal, which is the truncation (at zero) of a normally distributed random variable with zero mean and constant standard deviation, that is $u_i \sim N^+(0, \sigma_u)$.[22] The most important characteristic of this distribution is that the modal value of the inefficiency term (i.e. the most frequent value) is close to zero, and higher values of u_i are increasingly less likely (frequent). Stevenson (1980) relaxed the somehow strong assumption that the most probable value is being fully efficient by introducing the truncated normal distribution, which replaces the zero mean of the pretruncated normal distribution by a new parameter to be estimated. It should be pointed out that the identification of both random terms in these models relies on the one-sided nature of the distribution of u_i and not necessarily on the asymmetry of the inefficiency term (see Li 1996). In other words, if the inefficiency term could take both positive and negative values, it would not be distinguishable from the noise term.

Several comments are in order regarding the above distributional assumptions. First, all of them provide closed-form solutions for the distribution of the composed error term, making the direct application of ML straightforward. Newer models are appearing in the literature that do not yield tractable likelihood functions and must be estimated by simulated

[21]Both variances can also be estimated using the second and third moments of the composed error term taking advantage of the fact that, while the second moment provides information about both variances, the third moment only provides information about the asymmetric random conduct term.

[22]Note that, for notational ease, we use σ_u to indicate hereafter the standard deviation of the pretruncated normal distribution, and not the standard deviation of the post-truncated variable u_i.

maximum likelihood. See Parmeter and Kumbhakar (2014, Section 7) for an excellent review of recent contributions dealing with this issue. Second, rigidities derived from the nature of some inputs, regulations, transaction costs, information failures and other adjustment costs may prevent firms from moving instantly towards long-run optimal conditions. In this context, firms may not only find it optimal to remain inefficient in the short run, but also their inefficiency may persist from one period to the next. Two different approaches have been used in the literature to incorporate the dynamic nature of the decision-making process into efficiency analyses: reduced-form models that do not define explicitly a mathematical representation of dynamic behaviour of the firm, and structural models that make explicit assumptions regarding the objective of the firm and on a rule for forming expectations with respect to future input prices and technological advances. For a more comprehensive review of this literature, see Emvalomatis (2009).

Third, so far we have assumed that the inefficiency and noise terms are independently distributed. This could be a strong assumption for instance in agriculture, where noisy and seasonal fluctuations often affect productive decisions. The error components independence assumption has been recently relaxed by Bandyopadhyay and Das (2006) and Smith (2008). While the first authors assume that v_i and u_i are jointly distributed as normally truncated bivariate so that u_i is truncated below zero, the second uses the copula approach. The copula allows parameterizing the joint behaviour of v_i and u_i and tests the adequacy of the independence assumption.[23] The latter author also shows that the distribution of the composed error term can yield wrong skewness problems, making it difficult to estimate the above model by ML. The so-called wrong skew problem appears in a SFA model when the OLS residuals have skewness of the wrong sign relative to the SFA frontier model that includes a one-sided error term. In this case, ML estimation will almost always produce fully efficient scores as σ_u tends to zero. Also, the ML estimator is subject to significant biases when error component dependence is incorrectly ignored. Using a set of simulation exercises, Simar and Wilson (2010) show that the wrong skewness issue might also appear even when the underlying skewness is the correct one.

Finally, O'Donnell et al. (2010) show that the application of standard methods of efficiency analysis to data arising from production under uncertainty may give rise to spurious findings of efficiency differences between

[23]Other authors have used the copula method in other types of SFA applications. For instance, Amsler et al. (2014) employ them to model time dependence, while Carta and Steel (2012) suggest their use in modeling multi-output models.

firms. This may be a serious issue in many applications, such as agriculture, fishing or banking where production uncertainty is relatively high. To deal with this issue, Chambers and Quiggin (2000) found it convenient to treat uncertainty as a discrete random variable and proposed to model uncertainty in terms of a state contingent technology, where each state represents a particular uncertain event. They also show that all the tools of modern production theory, including cost and distance functions, may be applied to state contingent production technologies.[24]

Once the model has been estimated, the next step is to obtain the efficiency values for each firm. They are often estimated by decomposing the estimated residuals of the production function. Following Jondrow et al. (1982), both the mean and the mode of the conditional distribution of u_i given the composed error term ε_i can be used as a point estimate of u_i.[25] Three comments are in order regarding the point estimates of u_i. First, Wang and Schmidt (2009) show that inference on the validity of the chosen specification of the inefficiency term should not be carried out by simply comparing the observed distribution of $\hat{u}_i$ to the assumed distribution for u_i. To carry out this test, we should compare the distribution of $\hat{u}_i$ and $E(u_i|\varepsilon_i)$. In this sense, they propose nonparametric chi-square and Kolmogorov-Smirnov type statistics to perform this test properly. Second, the choice of a particular distribution for the inefficiency term should not only rely on statistical criteria, but also on the competitive conditions of the markets where the firms are operating. For instance, the above-mentioned distributions allow for the existence of very inefficient firms in the sample, which is an unappealing feature if they are operating in very competitive markets. For these markets, it might be more appropriate to use the double-bounded distribution introduced by Almanidis et al. (2010) that imposes both lower and upper theoretical bounds on the values of the inefficiency term. Moreover, the results of some recent papers providing evidence on the correlation between market power and operational inefficiency suggest that different market equilibrium outcomes might yield different distributions for the inefficiency term—e.g. Huang et al. (2017).

[24]Empirical application of the state contingent approach has proved difficult for several reasons because most of the data needed to estimate these models are lost in unrealized states of nature (i.e. outputs are typically observed only under one of the many possible states of nature).

[25]As aforementioned, firms' efficiency scores can also be computed without making specific distributional assumptions on the error components using the so-called *distribution-free approach*. As Kumbhakar et al. (2015, p. 49) remark, the drawback is that the statistical properties of the estimator of u_i may not be ready available.

Finally, the previous discussion is concerned with the technical side of the firm. Recent developments in duality theory allow the decomposition of overall economic efficiency into technical and allocative terms in a consistent way. For an updated presentation of the models based on numerical methods that are deterministic, we refer the reader to Aparicio and Zofío (2017). These authors show that, if the production or cost functions are *known*, the process is simplified when the analysis involves self-dual homogenous technologies, as in the Cobb-Douglas or generalized production function cases. When the estimation of unknown production or cost function is required, Parmeter and Kumbhakar (2014) summarize the existing methods, favouring those relying on the primal perspective that are easier to identify and estimate, over systems of equations based on the dual approach. Their preferred approach estimates a system consisting of a stochastic production function, which allows for technical inefficiency, and a set of first-order conditions for cost minimization. Departure from the optimality conditions can be measured by the difference in the bilateral ratios corresponding to the marginal productivities and input prices. A parametric decomposition of overall cost efficiency solves the system of equations by maximum likelihood for a given functional form (e.g. Translog). The error term of the stochastic production function can be decomposed using Jondrow et al. (1982) in order to compute firms' technical inefficiency, while the allocative inefficiencies are obtained from the residuals of the first-order conditions, and the input demands functions.[26] Again, the method could be adapted to decompose overall revenue and profit efficiency. For the latter case, see Kumbhakar et al. (2015), who also show how to implement these methods using the STATA software.

3.3 Evaluating, Comparing and Reconciling DEA and SFA Methods for Decision Making

Once the basic characteristics of the standard DEA and SFA approaches have been presented, it is clear that the individual results, rankings and distributions obtained from both methods will generally differ. However,

[26]However, Aparicio and Zofío (2017) show that the use of radial measures is inadequate to decompose cost efficiency in the case of nonhomothetic production functions because optimal input demands depend on the output targeted by the firm, as does the inequality between marginal rates of substitution and market prices—i.e. allocative inefficiency. They demonstrate that a correct definition of technical efficiency corresponds to the DDF.

the difference between the nonparametric and parametric methods is less pronounced nowadays than they used to be because both approaches now benefit from recent advances that address their shortcomings. In the past, it was their deterministic or stochastic nature and therefore their relative ability to accommodate noise and error that marked the difference. A second difference was their nonparametric and parametric nature, preventing second-order differentiability and proneness to misspecification error, respectively. In passing, we note that most DEA results are based on the envelopment programmes (10) and (11), which successfully identify reference peers, but do not offer the characterization of the production technology, except the nature of returns to scale. To unveil the characteristics of production technology and economic optima, one must resort to the "multiplier forms" (12) and (13) providing linear hyperplanes (facets), from which one gains information on the optimal weights and from there marginal productivities, rates of substitution and transformation, etc. Still, the mathematical programming approach does not enjoy the interpretative convenience of the econometric approach, e.g. in terms of average technical and economic elasticities (scale, cost, revenue, etc.). In turn, in this latter approach it is the "average" firm that characterizes technology, rather than the observations at the frontier, which are essentially those that represent "best practice" behaviour, i.e. those with the highest efficiency, productivity and optimal economic performance.

From the perspective of DEA, its deterministic nature has been qualified thanks to the extension of statistical methods to mathematical programming. This is the case of chance-constrained DEA and recent robust statistical techniques based on data resampling (bootstrapping, fuzzy methods, stochastic approaches, etc.), which can be customarily found in several software packages thanks to the increase in processing capacity—e.g. Álvarez et al. (2016). As for the need to adopt a specific—even if flexible—functional form in SFA, that may satisfy the desired regularity conditions locally, and be prone to misspecification bias, the availability of semiparametric and Bayesian techniques is opening new opportunities—e.g. Kumbhakar, Park, et al. (2007). Also, new proposals based on *Convex Nonparametric Least Squares* (CNLS) and the so-called *Stochastic Nonparametric Envelopment of Data* (StoNED) are also trying to bridge the gap between both methods—Johnson and Kuosmanen (2015).

The extent to which results obtained with both approaches differ is a general matter of concern that has been addressed by several authors, who employing the same data sets resort to nonparametric tests to compare the similarity of the distributions of the efficiency scores (see, e.g. Hjalmarsson et al. 1996 and Cummins and Zi 1998). Ultimately, what matters is the

ability to provide reliable results on individual performance, not only for the managers of the firms operating within an industry, but also for stakeholders and government agencies involved in regulation, competition and general policy analysis. Bauer et al. (1998) are worth referencing because they propose a set of consistency conditions for the efficiency estimates obtained using alternative methodologies. The consistency of results is related to: (1) the comparability of the estimates obtained across methods, assessed with respect to the efficiency levels (comparable means, standard deviations and other distributional properties), rankings and identification of best and worst firms; (2) the degree to which results are consistent with reality, determined in relation to their stability over time, in accordance with the competitive conditions in the market; and finally, (3) similarity with standard nonfrontier measures of performance (e.g. KPIs). In general, the higher the consistency of efficiency results across all these dimensions, the more confidence regulators and competition authorities will have on the conclusions derived from them and the intended effect of their policy decisions. These authors survey a number of studies on financial institutions and examine all these consistency conditions for regulatory usefulness.

Several authors have used meta-analyses to confront results from different studies and identify the main factors behind the observed variability in efficiency. For instance, Brons et al. (2005) and Bravo-Ureta et al. (2007) conclude that results from alternative models vary due to several factors including the methods used (e.g. nonparametric vs. parametric), alternative model specifications (e.g. returns to scale), specific observations and variables (e.g. nondiscretionary), time periods (e.g. cross-section or panel data), etc. Also, Odeck and Bråthen (2012) find that: (1) the random effects model outperforms the fixed effects model in explaining the variations in mean technical efficiencies; (2) studies relying on nonparametric DEA models yield higher values than SFA models (as expected given their deterministic nature); and (3) that panel data studies have lower scores as compared with those using cross-sectional data.

4 Dimension Reduction and Variable Selection

This section is devoted to the discussion of several issues related to firms' technology and the number of frontier determinants. Indeed, new technologies allow researchers to collect larger amounts of data. A relative trade-off exists between the number of observations and variables, serving

to determine the confidence level and reliability of results. This trade-off is summarized within the concept of degrees of freedom. Degrees of freedom is a function relating the sample size (I) with the number of independent variables (e.g. N inputs and M outputs), and the larger the number of independent observations with respect to the number of explanatory variables, the greater the confidence offered to researchers when making inferences about a statistical population (i.e. hypotheses testing). This serves for both DEA and SFA approaches when performing parametric tests relying on asymptotic theory, meaning that theoretical properties can be established for large samples (e.g. regarding parameter estimates, the nature of returns to scale, input and output substitutability, etc.).[27]

Moreover, besides the use of parametric tests, in DEA, the ability of these methods to discriminate observations by their efficiency is compromised when the number of observations is limited. As DEA methods search for the most favourable weights for the firm under evaluation, it is more likely to assign weights that render the firm efficient when there are less degrees of freedom. Any firm for which the ratio of outputs to inputs can be maximized by varying weights (including zero values) will be deemed efficient—i.e. any extreme firm employing the lowest amount of any of the N inputs or producing the largest amount of any of the M outputs is categorized as efficient *by default*. In a DEA context, this situation has resulted in a "rule of thumb" proposal by which the number of observations should be at least twice the number of inputs and outputs: $I \geq 2(N + M)$—Golany and Roll (1989), while Banker et al. (1989) raise this threshold to three. However, if the population is small with industries composed of just a few firms in a particular market, the DEA benchmarking results can still be helpful, while a regression based analysis may yield inconclusive results—regardless of the lack of statistical validity.

While the availability of massive data sets has reshaped statistical thinking, and computational power allows carrying out statistical analyses on large size databases, the trade-off between observations and variables persists as a relevant issue in many applications. Reducing the dimensions of data is a natural and sometimes necessary way of proceeding in an empirical analysis using either DEA or SFA. Indeed, dimension reduction and variable selection are the main approaches to avoid the curse of dimensionality.

[27]Alternative hypotheses testing methods corresponding to nonparametric and bootstrap-based inference have been proposed in the literature, see the chapter devoted to the statistical analysis of nonparametric benchmarking contributed by Leopold Simar, Camilla Mastromarco and Paul Wilson.

4.1 Dimension Reduction

This empirical strategy can be viewed as a two-stage procedure. In the first stage, a set of variables are aggregated into a small number of composites or aggregated variables. In a second stage, the composites are plugged into a production or economic frontier (e.g. profit) that is estimated using either nonparametric or parametric techniques. Therefore, this approach reduces the input (output) dimensionality of the data set by replacing a set of decision variables and regressors with a lower-dimensional function.

The most common methods used to achieve this objective are *Principal Component Analysis* (PCA) and *Explanatory Factor Analysis* (EFA). The dimensionality of the data set is reduced using these statistical methods by expressing the variance structure of the data through a weighted linear combination of the original variables. Each composite accounts for maximal variance while remaining uncorrelated with the preceding composite. These methods have been used to carry nonparametric efficiency analyses. For instance, Ueda and Hoshiai (1997) and Adler and Golany (2002) develop PCA-DEA models to obtain the efficiency estimates where a set of principal components replace the original variables. Other remarkable applications of this approach are Adler and Yazhemsky (2010) and Zhu (1998). As only a percentage of the information is retained from each of the original variables, the discriminatory power of the DEA model is improved. Yu et al. (2009), Nieswand et al. (2009), and Growitsch et al. (2012) use these PCA and EFA in a SFA framework to control for the effect of many environmental conditions on cost efficiency of electricity distribution networks.

From an analytical perspective, this two-stage procedure implicitly assumes that the technology is separable. Separability hinges on how the marginal rate of substitution between two individual variables only depends on the variables within the composite. Therefore, a necessary condition for the existence of a theoretically consistent composite is the separability of the elements within the aggregate from those outside the aggregate.[28] Otherwise, the use of composites in estimation may well be subject to specification errors. A tentative action is to test the existence of separability using cost and production functions as in Kim (1986). However, when the number of inputs (outputs) is large, the precision of these tests is probably

[28]A comprehensive discussion about the theoretical implications of different types of separability (e.g. strong vs. weak) can be found in Chambers (1988).

too low to be used with confidence. Moreover, carrying out such tests can be an impossible task when the dimensionality problem becomes truly severe.

Regardless the separability issue, PCA and EFA are *unsupervised* methods for reducing the dimension of the data in the terminology coined by Fisher (1922) because they only use information on how the input and output variables are statistically distributed, how large their variances are or whether they are highly correlated. That is, both methods ignore information on the dependent variable when reducing the dimension of the data. Therefore, their predictions might be biased because relevant predictive variables can be underweighted, while irrelevant factors can be overweighed. This type of error might explain the fact that clear relationships are not often obtained in many studies using PCA and EFA composites.

The so-called *supervised* or sufficient methods take somehow into account the relationship between the variable to be predicted and the vector of explanatory variables to be aggregated when they reduce the dimension of the data. See Bura and Yang (2011) for an overview of sufficient dimension reduction in regression. Li (1991) introduces the first method for sufficient dimension reduction, i.e. Slice Inverse Regression (SIR), and since then, various types of inverse regressions have been proposed. The inverse regression methods have been intensively applied in fields such as biology, genome sequence modelling and pattern recognition involving images or speech. However, the potential of sufficient dimension reduction methods for reducing data sets has barely been explored in economics. An exception is Naik et al. (2000) that use the SIR techniques to aggregate marketing data, and Orea et al. (2015) in the first attempt to apply supervised methods to production economics using a SFA model.

The popularity of inverse regression methods in other fields is due to the fact that most of them are computationally simple. We next describe briefly the procedure of the original SIR method. Simplifying the notation in Li (1991), the model to be estimated can be written as:

$$\ln y = \beta_0 + \beta_1 \ln f(\theta_1 x_1 + \theta_2 x_2) + \varepsilon, \tag{21}$$

where $\theta = (\theta_1, \theta_2)$ is a vector of unknown coefficients, and ε is a random term which is assumed to be independent of the inputs levels. This method makes no assumption about the distribution of the error term. This makes it appealing for SFA applications where the error term includes noise and inefficiency random terms. In this formulation, the response variable is related to x_1 and x_2 through the reduced 1-dimensional variable $X = \theta_1 x_1 + \theta_2 x_2$. SIR and other sufficient dimension reduction methods are developed to find the space generated by the unknown θ vector. This space should be

estimated from the data and is based on the spectral decomposition of a kernel matrix K that belongs to the central subspace (i.e. the smallest dimension reduction subspace that provides the greatest dimension reduction in the predictor vector). To this aim, Li (1991) proposes to reverse the conventional viewpoint in which y is regressed on $\boldsymbol{X}$ and showed that a PCA on a nonparametric estimate of $\boldsymbol{E}(\boldsymbol{X}|y)$ can be used to estimate K. The approach relies on partitioning the whole data set into several slices according to the y-values. Thus, the dependent variable is only used to form slices while the standard PCA does not use any information from y. As the above procedure provides a rather crude estimate of $\boldsymbol{E}(\boldsymbol{X}|y)$, other first-moment based methods have been proposed. For instance, parametric inverse regression aims to estimate the central subspace using least squares. This parametric version of inverse regression regresses each (standardized) input variable on a set of arbitrary functions of y. The fitted values of $\boldsymbol{E}(\boldsymbol{X}|y)$ are then used as an estimate of K.

Regarding the choice between supervised and unsupervised variable dimension reduction methods, it is apparent from the above discussion that we will always get better results using a supervised method than an unsupervised one. In theory, this is true, but not in practice. Note that the supervised methods need to control the relationship between the variable to be predicted and the vector of explanatory variables when they proceed with the aggregation. In practice, this relies on a PCA of a (non)parametric estimate of $\boldsymbol{E}(\boldsymbol{X}|y)$. In this sense, Adragni and Cook (2009) point out that some of the best moment-based methods turned out to be rather inefficient in relatively simple settings. Thus, in any particular efficiency analysis it could occur that $\boldsymbol{E}(\boldsymbol{X}|y)$ is too poorly estimated meaning that an unsupervised method might yield better results. In order to minimize biases associated with inaccurate inverse regressions, we thus suggest using more recent, albeit more complex supervised methods.[29] Even then, a limitation of these variable dimension reduction methods in efficiency analysis is that their sample property results have been obtained with normally distributed error components. A field of future research is the analysis of their properties in SFA models with asymmetric error terms.

A final issue that should be examined in this subsection is determining the number of composites to retain. To choose the number of composites,

[29]For instance, whereas Xia et al. (2002) and Bura (2003) propose semiparametric techniques to estimate the inverse mean function, $E(X|Y)$, Cook and Ni (2005) develop a family of dimension reduction methods by minimizing Quadratic discrepancy functions and derive the optimal member of this family, the inverse regression estimator.

we propose using conventional model selection tests that balance the lack of fit (too few composites) and overfitting (too many composites). The use of model selection tests is usually restricted to cases where economic theory provides no guidance on selecting the appropriate model, and the alternative models are not nested, as in the present approach.

4.2 Variable Selection

In biology, industrial engineering and other noneconomic fields, the elimination of variables is highly desirable as they are mainly interested in predicting a response (dependent) variable, and the cost of overfitting (i.e. estimating a more complex model than it needs to be) is the increased variability of the estimators of the regression coefficients. In economics, we should at least add three additional reasons to proceed with the elimination of variables. First, the "dimensionality" issue becomes acute when flexible functional forms are estimated as the number of parameters increases more rapidly when interactions are considered, or the semiparametric or nonparametric techniques require a manageable number of explanatory variables to be implemented. Second, for interpretability, the identification of relevant variables based on economic theory or "expert" knowledge may or may not be correct if the model is overfitted. Finally, it is always preferable to build a parsimonious model for easier data collection. This is especially relevant in efficiency analyses in regulated industries where the regulators need to collect costly data on a large set of variables in order to control for many geographical, climatic or network characteristics of the utilities sector that affect production costs, but which go unobserved.

Many different procedures and criteria for selecting the best regression model have been suggested. See Mittelhammer et al. (2000) for a general and critical analysis of the variable selection problem and model choice in applied econometrics. The so-called backward, forward and stepwise procedures may lead to interpretable models. However, the results can be erratic as any single test used at some stage in the above mentioned procedures is not indicative of the operating characteristics of the joint test represented by the intersection of all the individual tests used. That is, because the subset selection is a discrete process, small changes in the data can lead to very different statistical models, and the sampling properties of these processes are virtually unknown. In addition to these procedures, other criteria such as Akaike's criterion (AIC), Schwarz's Bayesian criterion (SBC) and some of their variants have been used to evaluate the impact of adding/removing

variables in regression models. For more details about these criteria and the associated penalty functions, see Fonseca and Cardoso (2007). The penalty term penalizes very complex models and increases with the number of parameters of the model. Thus, these criteria involve minimizing an index that balances the lack of fit (too few variables) and overfitting (too many variables).

It should be pointed out that dimensionality reduction and variable selection also carry a long tradition in the DEA literature. There are methods that study correlation among variables, with the goal of choosing a set that do not represent largely associated values. However, these approaches may yield unreliable results because the removal of even highly correlated variables can still have a large effect on the DEA results—Nunamaker (1985). For instance, Lewin et al. (1982) and Jenkins and Anderson (2003) apply regression and multivariate analysis to select and reduce the number of variables in the DEA model, respectively. In both cases, they study the explanatory power of the variables using a stepwise method by which variables are included sequentially. The latter regression yields a better goodness of fit, and based on these results, they select the specific inputs, outputs and nondiscretionary variables to be included in the DEA model. It is interesting to note that Dyson et al. (2001), when studying several pitfalls and protocols in DEA, call for exercising caution when simply dropping some variables based on their high correlation (e.g. inputs) since reference hyperplanes, and therefore efficiency scores, can change significantly. As the sequential regression method suggested by Lewin et al. (1982) is influenced by the collinearity between regressors, it is prone to the previous selection problem.

To precisely address the arbitrariness and problems related to discarding variables based on their high correlations with those ultimately retained in the analysis, Jenkins and Anderson (2003) propose a multivariate analysis to reduce the dimensionality of variables. After reviewing previous proposals based on multivariate analysis, such as canonical correlation (Sengupta 1990; Friedman and Sinuany-Sterns 1997), discriminant analysis (Sinuany-Sterns and Friedman 1998) and the already mentioned principal components analysis (Ueda and Hoshiai 1997; Adler and Golany 2002), these authors propose a multivariate method to identify which variables can be discarded with least loss of information. The method is based on the variance of the input or output about its mean value, for if its value is constant, then it plays no part in distinguishing one DMU from another. On the contrary, a large variation indicates an important effect. Comparing the results obtained with their method using the databases of several published studies

confirms the worries expressed by Dyson et al. (2001) as there are large variations in the computed efficiencies.

A second strand of literature examines whether removing or adding variables in a sequential way results in significant changes in the DEA efficiency distributions—Norman and Stoker (1991). In this approach, variables are to be discarded or included according to a selection process that assesses their statistical significance. In this vein, Kittelsen (1993) surveys several tests to establish the significance of change in efficiency results when sequentially removing or adding variables. He shows that the usual tests (*F*, Kolmogorov–Smirnov, *t*-ratio, etc.) used to determine if the subsequent efficiency distributions remain the same or change after removing or adding variables are invalid because they assume that the scores are independently and identically distributed. This is not the case with the sequential method because the individual DEA efficiencies are nested, with those corresponding to the augmented model including more variables (restrictions) and presenting larger scores—a property deriving from linear optimization theory and resulting also in more efficient firms.

This is a valuable exercise that nevertheless should be revisited with the use of tests that take into account the nested nature and distributional forms of the efficiency scores as proposed by Pastor et al. (2002). These authors define a new "efficiency contribution measure" (ECM, representing the marginal impact on efficiency of a variable) that compares the efficiency scores of two radial DEA models differing in one output or input variable (termed candidate). Then, based on this ECM, at a second stage a statistical test is developed that allows an evaluation of the significance of the observed efficiency contribution of the differing variable (i.e. the effects above a level of tolerance or threshold). This test provides useful insights for the purpose of deciding whether to incorporate or delete a variable into/from a given DEA model, on the basis of the information supplied by the data. Two procedures for progressive selection of variables are designed by sequentially applying the test: a forward selection and a backward elimination. This method for selecting the variables to be included in the model is reworked by Wagner and Shimshak (2007), who improve these procedures by formalizing a stepwise method which shows how managers can benefit from it in a structured decision making process.

We now refer to variable selection procedures involving hundreds or thousands of explanatory variables, i.e. model selection with high dimensional data sets. The traditional methods face significant challenges when the number of variables is comparable to or larger than the sample size. These challenges include how to make the estimated models interpretable, in our

case, from an economic perspective. An approach to proceed with variable selection with high dimensional data sets is *Penalized Least Squares* (PLS), which is a method that tends to produce some coefficients that are exactly zero. As this outcome is equivalent to a reduction in candidate explanatory variables from the model, LASSO and other PLS estimators help in getting more interpretable models.

As remarked by Fan and Lv (2010), what makes high dimensional statistical inference possible is the assumption that the *underlying* (distance) function does have less variables than the data set. In such cases, the *d*-dimensional regression *parameters* are assumed to be sparse with many components being zero, where nonzero components indicate the important variables. With sparsity, variable selection can improve the estimation accuracy by effectively identifying the subset of important predictors and can enhance the model interpretability with parsimonious representation. Many variable selection criteria or procedures are closely related to minimizing the following PLS:

$$\frac{1}{2I}\sum_{i=1}^{I}(y_i - x_i\beta)^2 + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \tag{22}$$

where d is the dimension of x_i, and $p_{\lambda_j}(\cdot)$ is a penalty function indexed by the regularization parameter $\lambda \geq 0$, controlling for model complexity. The dependence of the penalty function on j is very convenient in production and cost analyses as it allows us to keep certain important explanatory variables in the model (e.g. key inputs in a production function or the output and input prices variables in a cost function) thus choosing not to penalize their coefficients. The form of the penalty function determines the general behaviour of the estimator. With the entropy or L_0-penalty, the PLS in (22) becomes

$$\frac{1}{2I}\sum_{i=1}^{I}(y_i - x_i\beta)^2 + \sum_{j=1}^{d} \lambda^2|M|, \tag{23}$$

where $|M|$ is the size of the candidate model. In this formulation, among models with the same number of variables, the selected model is the one with the minimum residual sum of squares. With the L_1-penalty specifically, the PLS estimator is called LASSO in Tibshirani (1996). When $p \leq 1$, the PLS automatically performs variable selection by removing predictors with very small estimated coefficients. The LASSO estimator satisfies the sparsity condition as it should automatically set small estimated coefficients to zero

in order to accomplish variable selection. However, it is a biased estimator, especially when the underlying coefficient of dropped variables is large.

The PLS approach can also be easily extended to the likelihood framework. Define a penalized likelihood function as:

$$Q(\beta) = \frac{1}{I}\sum_{i=1}^{I} \ln LF(y_i, x_i; \beta) - \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|). \tag{24}$$

Maximizing the penalized likelihood results in a penalized estimator. For certain penalties, the selected model based on the penalized likelihood satisfies $\beta_j = 0$ for specific β_j's. Therefore, parameter estimation is performed at the same time as the model selection. As the likelihood framework is the most used framework in the SFA literature, we believe that this is a promising area of research for the near future when we will progressively be able to collect larger data sets.

4.3 The Choice Between Variable Dimension Reduction and Variable Selection

We conclude this section with a practical discussion concerning the choice between variable dimension reduction and variable selection in efficiency analyses. Variable dimension reduction is appealing when: (1) the main issue is the overall effect of a wide-ranging (holistic) phenomenon formed by a large number of factors with complex interactions, and not the partial effect of its components; and (2) it is difficult to either formulate hypotheses associated with these variables or impose restrictions derived from production theory on the technology. In this sense, environmental variables are good candidates for aggregation using some of the techniques outlined above. On the other hand, this approach is probably more suitable in DEA applications where researchers' main interest is in measuring firms' inefficiency and performing benchmarking exercises, and not in disentangling *specific* technological characteristics such as economies of scale, scope and substitution between inputs and outputs.

5 Controlling for Observed Environmental Conditions

The concern about the inclusion of environmental variables (also called *contextual*, nondiscretionary or *z*-variables) has generated the development of several models either using parametric, nonparametric or semiparametric

techniques. Although we do not pretend to provide a complete survey of the alternatives for including *z*-variables, given the wide range of models that have been developed, here we only mention the methods most frequently applied. For a more detailed review of this topic in SFA, see Parmeter and Kumbhakar (2014). A brief summary of this issue in the nonparametric literature can be found in Johnson and Kuosmanen (2012).[30]

5.1 Frontier Determinants vs. Determinants of Firms' Inefficiency

The first methodological choice is whether we should incorporate the *z*-variables as either frontier determinants, determinants of firms' inefficiency or as determinants of both the frontier and the inefficiency term. While the above dilemma may not be very relevant in practice as the sign of the contextual variables is not necessarily assumed to be known beforehand, the key question that should be responded in order to include the *z*-variables as frontier determinants is whether a fully efficient firm will need to use more inputs to provide the same services or produce the same output level if an increase in a contextual variable represents a deterioration in the environment where it operates. To respond properly to this question most likely requires having a good knowledge of the industry that is being examined or recurring to technical (e.g. engineering) support. In general, we should include as frontier drivers those variables that are fundamental to production.

Whether *z*-variables should be included in the frontier function or the inefficiency term may *not* be a semantic issue from a conceptual point of view and might have very different implications for policy makers, regulators and managers. For instance, the traditional time trend can be viewed as a noncontrollable *z*-variable. If this variable is added to the production or cost frontier, it captures technical change. A poor rate of technical progress might suggest implementing policy measures encouraging R&D activities. In contrast, if the same variable is included as a determinant of firms' inefficiency, it captures changes in firms' inefficiency over time. In this case, deterioration in firms' performance might suggest implementing policy measures aiming to improve (update) managerial skills.

[30]The chapter by John Ruggiero discusses environmental variables and how to render observations comparable in performance studies.

The above distinction may also be important in regulated industries where regulators purge firms' cost data in order to control for differences in environmental conditions. In these settings, can firms use unfavourable weather conditions as an excuse to avoid being penalized due to their bad performance? As the environment is not controlled by the firm, one might argue that firms should not be blamed for environment-induced inefficiency. This interpretation implies that regulators should purge firms' cost data when environmental conditions have both direct and indirect effects on firms' cost. We should remark, however, that purging the data completely is likely to be a fairer policy in the short run, i.e. conditional on current firms' managerial skills. However, if the estimated indirect effect is significant, one could conclude that not compensating all weather effects could help to encourage these firms to hire better qualified executives and staff, perhaps not immediately, but at least in the long run. Thus, regulators might be aware of this trade-off between short-run and long-run objectives when they design their incentive schemes.

5.2 DEA Estimation of the Effects of Contextual Variables

The inclusion of environmental variables in DEA has been done in one or two stages.[31] The one-stage DEA approach (hereafter 1-DEA) is to augment the model by treating the z-variables as inputs or outputs that contribute to defining the frontier. For instance, the DDF *with* environmental variables would be:

$$D_V\left(y,x,z;-g_x,g_y\right)=\max\left\{\beta:(x-\beta g_x,y+\beta g_y)\in V(z)\right\}. \quad (25)$$

In the two-stage DEA method (hereafter 2-DEA), the efficient frontier and the firm-level efficiency scores are first estimated by DEA or other nonparametric method using a representation of the firm's technology *without* environmental variables, as in (1). Let $\hat{E}_i$ denote the first-stage estimate of firm's (x_i, y_i) efficiency level. In the second stage, the estimated DEA efficiency scores are regressed on contextual variables. The two-stage regression can be written in general terms as:

[31]Although the two-stage method is the most popular one in DEA for identifying inefficiency determinants, three-stage models have also been developed (see, e.g., Fried et al. 2002).

$$\hat{E}_i = \tau z_i + \varepsilon_i \geq 1, \tag{26}$$

where τ is a vector of parameters, and ε_i is a random variable. The inequality in (26) yields a truncated (linear) regression model.[32] From the equations above, it is straightforward to notice that while the one-stage methods incorporate the *z*-variables as frontier determinants, the two-stage methods incorporate them as determinants of firms' inefficiency, which in turn is measured with respect to an uncorrected production (or cost) frontier. This difference implies that the sign of the contextual variables is assumed to be known beforehand in one-stage DEA methods, whereas the sign of these variables is estimated in two-stage methods. Thus, from a conceptual point of view, 2-DEA methods are more appropriate in applications where the environment is multifaceted and consists of a large and varied number of factors with complex interactions, so that it is difficult to formulate hypotheses with respect to the effect of weather conditions on firms' performance.

The choice of a proper method to control for environmental conditions has attracted merited attention in the DEA literature. The seminal paper of Banker and Morey (1986) modified the measure of inefficiency obtained by removing the effect of *contextual* variables on the measured inefficiency level within the DEA model. Ruggiero (1996) and other authors have highlighted that the one-stage model introduced by the previous authors might overestimate the level of technical inefficiency. To solve this problem, other models using several stages have been developed. Ray (1988) was the first to propose a second stage where standard DEA efficiency scores were regressed on a set of contextual variables. The 2-DEA method was widespread until Simar and Wilson (2007) demonstrated that this procedure is inconsistent because it lacks a coherent data generating process and the first-stage DEA efficiency estimates are serially correlated. The problems arise from the fact that (26) is the assumed model, whereas the true model is $E_i = \tau z_i + \varepsilon_i \geq 1$. Here, the dependent variable is unobserved and must be replaced by an estimate $\hat{E}_i$. Simar and Wilson (2007) show that, unfortunately, $\hat{E}_i$ is a biased estimator of E_i because, by construction, z_i is correlated with the error term ε_i. To address this issue, these authors propose the use of a bootstrap method to correct for the small sample bias and serial correlation of the DEA efficiency

[32]Interesting enough, this specification of the way efficiency scores depend on *z*-variables corresponds to the popular KGMHLBC model in the SFA approach (see next subsection).

estimates. Further, they advocate the use of the truncated regression model that takes into account explicitly the bounded domain of the DEA efficiency estimates.[33]

Since this remarkable paper, the statistical foundations of the 2-DEA method have been subject to intensive debate. For instance, Banker and Natarajan (2008) show that the second-stage OLS estimator of the contextual variables is statistically consistent under certain assumptions and regularity conditions. Subsequent discussion has focused on the assumptions. Banker and Natarajan (2008) argue that their statistical model allows for weaker assumptions than the model of Simar and Wilson (2007). In turn, Simar and Wilson (2010) discuss the assumptions made by Banker and Natarajan and find them rather restrictive. On these grounds, Johnson and Kuosmanen (2012) further elaborate the assumptions and the statistical properties of the two-stage estimators under more general assumptions.

These latter authors also develop a new one-stage semi-nonparametric DEA-style estimator that facilitates joint estimation of the frontier and the effects of contextual variables. They introduce the contextual variables to the already mentioned StoNED model, where the z-variables are incorporated additively to the parametric part of the model, which is estimated jointly with the nonparametric frontier. The new StoNED method is similar to the 1-DEA in that it jointly estimates the frontier and the contextual variables using convex nonparametric least squares regression. Both models mainly differ in the assumption made with respect to the truncated noise term. In the recently developed semiparametric literature, it is worthwhile mentioning another two models that also allow controlling for environmental variables. The first one is the *Semiparametric Smooth Coefficient Model* (SPSCM) introduced by Li et al. (2002) where the regression coefficients are unknown functions, which depend on a set of contextual variables. Sun and Kumbhakar (2013) extend this model by allowing the environmental variables to also enter through the inefficiency. The second model is the *Latent Class Model* (LCM), where z-variables enter in nonlinear form for the probabilities of belonging to the classes (see, e.g., Orea and Kumbhakar 2004).

[33]Daraio and Simar (2005) propose an alternative approach by defining a conditional efficiency measure. This approach does not require a separability condition as demanded by the two-stage approach.

5.3 The Inclusion of Contextual Variables in SFA

Like the two-stage DEA method, early papers aiming to understand firms' inefficiency using the SFA approach proceeded in two steps. In the first step, one estimates the stochastic frontier model and the firms' efficiency levels, ignoring the z-variables. In the second step, one tries to see how efficiency levels vary with z. It has long been recognized that such a two-step procedure will give biased results (see, e.g., Wang and Schmidt 2002). The solution to this bias is a one-step procedure based on the correctly specified model for the distribution of y given x and z.

Once we have decided to treat the z-variables as inefficiency determinants and hence heteroscedastic SFA models are to be estimated, a second methodological choice appears: How to do it. Summaries of this literature can be found in Kumbhakar and Lovell (2000) and Parmeter and Kumbhakar (2014). The available options can be discussed using the general model introduced by Álvarez et al. (2006)[34]:

$$y_i = x_i'\beta + v_i - u_i, \tag{27}$$

where $x_i'\beta$ is the log of the frontier production (distance) function (e.g. Translog), $u_i \sim N^+\left(\mu_i, \sigma_{ui}^2\right)$, $\mu_i = \exp\left(\delta_0 + z_i'\delta\right)$, $\sigma_{ui} = \exp\left(\gamma_0 + z_i'\gamma\right)$, and δ_0, δ, γ_0 and γ are parameters to be estimated, and z_i is the vector of efficiency determinants. The environmental variables enter into the model through both the pretruncated mean and variance of the inefficiency term, and hence, the model allows for nonmonotonic effects of the z-variables on firms' inefficiency. According to this model, Álvarez et al. (2006) divide most heteroscedastic SFA models into three groups. In the KGMHLBC-type models, it is assumed that the variance of the pretruncated normal variable is homoscedastic (i.e. $\gamma = 0$) and, thus, the contextual variables are introduced here through the pretruncated mean. In contrast, in the RSCFG-type models, it is assumed that the mean of the pre-truncated normal variable is homoscedastic (i.e. $\delta = 0$) and, hence, the environmental variables are treated as determinants of the variance of the pretruncated normal variable. Finally, the contextual variables are introduced in the general models through both the mean and the variance of the normal distributed random variable. Some of the above models satisfy the so-called *scaling property* in the

[34]The general models introduced by Wang (2002) and Lai and Huang (2010) are similar, but they parameterize the pretruncation mean of the distribution as a linear function of the z-variables.

sense that the inefficiency term can be written as a deterministic (scaling) function of a set of efficiency covariates times a one-sided random variable that does not depend on any efficiency determinant. That is,

$$u_i = h_i(z_i, \gamma)u_i^*, \quad u_i^* \sim N^+\left(\mu, \sigma_u^2\right). \tag{28}$$

As Parmeter and Kumbhakar (2014) point out, the ability to reflect the scaling property requires that both the mean and the variance of the truncated normal are parameterized identically and with the same parameters in each parameterization. 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. That is, the scaling property implies that changes in z_i affect the scale but not the shape of u_i. In this model, u_i^* can be viewed as a measure of basic inefficiency which captures things like the managers' natural skills, which we view as random. How well these natural skills are exploited to manage the firm efficiently depends on other variables z_i, which might include the manager's education or experience, or measures of the environment in which the firm operates.

Although it is an empirical question whether or not the scaling property should be imposed, it has some features that make it attractive to some authors (see, e.g., Wang and Schmidt 2002). Several authors have found the scaling property useful to remove individual fixed effects and still get a closed form for the likelihood function (Wang and Ho 2010), to address endogeneity issues (Griffiths and Hajargasht 2016), to relax the zero rebound effect assumption in traditional demand frontier models (Orea et al. 2015), or to allow for spatial correlation among firms' efficiency (Orea and Álvarez 2019). From an statistical point of view, the most fundamental benefit of the scaling property is that the stochastic frontier and the deterministic component of inefficiency can be recovered without requiring a specific distributional assumption on u_i. Indeed, if we take into account our specification of firms' inefficiency in (28) and define $u^* = E(u_i^*)$, then taking expectations in (27) yields:

$$y_i = x_i'\beta - h_i(z_i, \gamma)\, u^* + \varepsilon_i^*, \tag{29}$$

where $\varepsilon_i^* = v_i - h_i(z_i, \gamma)\left(u_i^* - u^*\right)$. Equation (29) can be estimated by *Nonlinear Least Squares* (NLLS).[35]

[35]Parmeter and Kumbhakar (2014) show that, if z_i and x_i do not include common elements, the conditional mean $E[u_i|z_i]$ can be estimated in a nonparametric fashion without requiring distributional assumptions for u_i.

6 Endogeneity Issues and the Choice of Orientation

Endogeneity problems can arise in stochastic frontier models if the frontier determinants are correlated with the noise term, the inefficiency term or both. As noted by Kumbhakar et al. (2013), the endogeneity issue is typical in econometric models, especially when economic behaviours are believed to affect both inputs and/or outputs levels (Kumbhakar and Tsionas 2011) and inputs and/or outputs ratios (Tsionas et al. 2015). On the other hand, in cost (profit) settings, endogeneity problems might appear when the outputs' levels (prices) or input prices depend on random shocks and economic inefficiency. This might happen if firms are allocative inefficient, or firms have market power as, in this case, input/output prices are not set competitively in the market. Although endogeneity issues were first discussed in the regression framework, it has also been addressed in the programming approach. Therefore, in this section we present a series of models addressing endogeneity in the nonparametric DEA and parametric SFA frameworks.

6.1 Endogeneity in DEA Models

It should be mentioned that ignoring the random nature of the data generating process does not preclude the existence of endogeneity problems in the calculation of efficiency in DEA models when the regressors (inputs or outputs) are correlated with technical inefficiency. Initially, Wilson (2003) surveys a number of tests that can be used to determine the independence and uncorrelated hypotheses in the context of efficiency measurement, including their merits and drawbacks. Afterwards, this author performs Monte Carlo simulations to establish that these tests have poor size properties and low power in moderate sample sizes. Peyrache and Coelli (2009) build upon these previous findings and propose a semiparametric Hausman-type asymptotic test for linear independence (uncorrelation). Resorting to Monte Carlo experimentation they show that it has good size and power properties in finite samples. Additionally, Cordero et al. (2015) show that with low and moderate levels of correlation, the standard DEA model performs well, but that for high levels—either negative or positive, it is mandatory to use instrumental techniques that correct the bias.

Based on these findings, Santín and Sicilia (2017) devise a semiparametric strategy similar to the instrumental variables approach in regression analysis, and that results in a DEA specification that accounts for the exogenous part of the

endogenous regression and that is uncorrelated with technical efficiency. As in the parametric counterpart, the first step is to choose an instrumental variable z that is significantly correlated with the endogenous regressor x (relevance), but uncorrelated with the true efficiency (exogeneity). Empirically, the relevance condition can be tested regressing the endogenous regressor x on the exogenous regressors and the instrument. As for the second condition of exogeneity, it cannot be tested since it is unobserved. In that case, these authors suggest interpreting it as the absence of correlation between the instrument z and the variables characterizing the alternative dimension of the production process, e.g. outputs y in the case of endogenous inputs. In that case, z should not have a partial effect on y (beyond its effect on the endogenous input) and should be uncorrelated with any other omitted variables (when this is the cause of endogeneity). Under these assumptions, the authors implement an instrumental variable process by substituting the estimated exogenous regressor for the endogenous regressor when solving the DEA programme associated with the relevant orientation, e.g. output. They find using Monte Carlo experiments that both standard and instrumental DEAs yield similar results in the case of low correlation, but that the latter clearly outperforms the former under high correlation. Also, coinciding with Orme and Smith (1996) the instrumental DEA performs better as the sample size increases.

6.2 Endogeneity in SFA Models

Researchers need to deal with endogeneity issues because the usual procedures for estimating SFA models depend on the assumption that the inputs are exogenous. However, dealing with the endogeneity issue is relatively more complicated in a SFA framework than in standard regression models due to the special nature of the error term. Several authors have recently proposed alternative empirical strategies to account for endogenous regressors in SFA settings. Some of them allow only for correlations between the regressors and the noise term, while other authors allow for correlations with the inefficiency term. Models can be estimated using *Instrumental Variables* (IV) techniques, *Maximum Likelihood* (ML) procedures or *Bayesian* estimation methods. Moreover, many of them can be estimated in one or two stages. Therefore, the researcher has several methods at hand to deal with endogeneity issues when estimating a SFA model. In the next paragraphs, we outline the main features of these methods, trying to identify their relative advantages and disadvantages.

Let us first assume that we are interested in estimating the following production model with endogenous regressors and panel data:

$$\begin{aligned} \ln y_i &= x_i'\beta + v_i - u_i, \\ x_i &= z_i'\delta + \eta_i, \end{aligned} \tag{30}$$

where x_i is a vector of endogenous variables (excluding $\ln y_i$), and z_i is a vector of exogenous or instrumental variables, and the second equation in (30) can be viewed as a reduced-form expression that links the endogenous variables with the set of instruments. The endogeneity problem arises if η_i in the second equation is correlated with either v_i or u_i in the first equation.

In order to estimate consistently the frontier model (30), Guan et al. (2009) propose a two-step estimation strategy. In the first step, they ignore the structure of the composed error term and suggest estimating the frontier parameters using a *Generalized Method-of-Moments* (GMM) estimator as long as valid instruments are found. In the second step, distributional assumptions are invoked to obtain ML estimates of the parameter(s) describing the variance of v_i and u_i, conditional on the first-stage estimated parameters. If the inefficiency term follows a homoscedastic distribution, only the GMM intercept is biased. However, we should be aware that ignoring that in the first stage of the process, the inefficiency term depends on a set of covariates could bias *all* model parameters. Indeed, a relevant issue that is often ignored when using OLS or GMM in a stochastic frontier framework is the endogeneity problem caused by the so-called left-out variables (Wang and Schmidt 2002), which arises because variables influencing technical inefficiency are ignored when estimating the model. Guan et al. (2009) mention this issue, but do not discuss its implications for the GMM estimation. To achieve consistent estimates, it is critical to ensure that the chosen instruments do not include determinants of u_i.

In line with the current chapter, Kumbhakar et al. (2013) and Malikov et al. (2015) suggest bringing economic behaviour into the analysis to solve endogeneity problems. Instead of introducing instruments for these endogenous variables in an ad hoc fashion (e.g. temporal lags of inputs and outputs), they address the endogeneity issue by defining a system in which they bring additional equations for the endogenous variables from the first-order conditions of profitability (cost) maximization (minimization). They advocate using a system approach for two reasons. First, estimates of allocative inefficiencies can be obtained from the residuals of the first-order conditions. Second, since the first-order conditions contain the same technological parameters, their estimates are likely to be more precise (efficient). However, estimation of such a system requires availability of input and output prices.

Their identification strategy also relies on competitively determined output and input prices as a source of exogenous variation.[36]

Other authors make efforts to address the endogeneity problem in a fully maximum likelihood estimation context. They use likelihood based instrumental variable estimation methods that rely on the joint distribution of the stochastic frontier and the associated reduced-form equations in (30). The simultaneous specification of both types of equations has the advantage that it provides more efficient estimates of the frontier parameters as well as improvement in predicting the inefficiency term. For instance, Kutlu (2010) proposes a model that aims to solve the endogeneity problem due to the correlation between the regressors and the two-sided error term.[37] He assumes that the error terms in (30) satisfy the following:

$$\begin{pmatrix} \Omega_\eta^{-1/2}\eta_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{I}_p & \rho\sigma_v \\ \rho'\sigma_v & \sigma_v^2 \end{pmatrix} \right) \tag{31}$$

where Ω_η is the variance-covariance matrix of η_i and ρ is a correlation vector between v_i and η_i. Based on (31), the equations in (30) can be written as:

$$\ln\, y_i = x_i\beta + \tau(x_i - z_i\delta) + \omega_i - u_i, \tag{32}$$

where $\omega_i = (1 - \rho'\rho)v_i$ and $\tau = \sigma_v\rho'\Omega_\eta^{-1/2}$, which can be viewed as a correction term for bias. Note that $\omega_i - u_i$ is conditionally independent from the regressors given x_i and z_i. Hence, conditional on x_i and z_i, the distribution of the composed error term in (32) is exactly the same as their traditional counterparts from the stochastic frontier literature. They then show that for the sample observations (y_i, x_i, z_i), the joint log-likelihood function of y_i and x_i is given by

$$\ln\, L(\theta) = \ln\, L_{y|x}(\theta) + \ln\, L_x(\theta), \tag{33}$$

where

$$\begin{aligned} \ln L_{y|x}(\theta) = &-\frac{I}{2}\ln(\sigma_\omega^2 + \sigma_u^2) \\ &+ \frac{1}{(\sigma_\omega^2 + \sigma_u^2)^{1/2}} \sum_{i=1}^{I} \ln \Phi\left(-(\ln y_i - x_i\beta - \tau(x_i - z_i\delta))\sigma_u/\sigma_\omega\right) \\ &- \frac{1}{2(\sigma_\omega^2 + \sigma_u^2)} \sum_{i=1}^{I} (\ln y_i - x_i\beta - \tau(x_i - z_i\delta))^2, \end{aligned} \tag{34}$$

[36]Kumbhakar (2011) also relies on profitability maximization, but he solves the endogeneity of both outputs and inputs first by deriving a particular form of the estimating equation in which the regressors are ratios of inputs and outputs. Thus, his transformed specification can be estimated consistently by ML methods using standard stochastic frontier software.

[37]In his model, the distribution of u_i is not allowed to have efficiency determinants.

and

$$\ln L_x(\theta) = -\frac{I}{2}\ln\left(\left|\Omega_\eta\right|\right) - \frac{1}{2}\sum_{i=1}^{I}\eta_i'\Omega_\eta^{-1}\eta_i. \tag{35}$$

The first part of the log-likelihood function (33) is almost the same as that of a traditional stochastic frontier model where the residual is adjusted by the $\tau(x_i - z_i\delta)$ factor. The second part is just the likelihood function of a multivariate normal variable. The likelihood function (33) can be maximized to obtain consistent estimates of all parameters of the model. However, if computational difficulties appear, one can use a two-step maximum likelihood estimation method. In the first stage, $\ln L_x(\theta)$ is maximized with respect to the relevant parameters. In the second stage, conditional on the parameters estimated in the first stage, $\ln L_{y|x}(\theta)$ is maximized. However, the standard errors from this two-stage method are inconsistent because the estimates are conditional on the estimated error terms from the first stage. Kutlu (2010) suggests using a bootstrapping procedure in order to get the correct standard errors. Alternatively, an analytical approach is possible as remarked by Amsler et al. (2016, p. 284).

The above mentioned ML model does not address the potential correlation with the inefficiency term, and neither does it assure consistency of parameter estimates when η_i is correlated with both v_i and u_i. Amsler et al. (2016) is the first paper to allow endogeneity of the inputs with respect to statistical noise and inefficiency separately. They propose using a copula in order to specify the joint distribution of these three random variables.[38] They select a multivariate normal (or "Gaussian") copula that does not permit to analytically integrate u_i out from the joint density for v_i, and η_i. For this reason, the parameter estimates should be obtained by maximum simulated likelihood estimation, where the joint density is approximated by taking many draws from the distribution of u_i and averaging.[39] One obvious difficulty with this approach is the need to specify a copula. Another difficulty of this approach is that it may be computationally challenging. Tran and Tsionas (2015) also use a Gaussian copula function to directly model the dependency of the endogenous regressors and the composed error

[38]A copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform.

[39]Applications of simulations to evaluate a likelihood can be found in Greene (2005, p. 24), Amsler et al. (2016), and Parmeter and Kumbhakar (2014; Sects. 6 and 7).

without using instrumental variables. Consistent estimates can be obtained by maximizing the likelihood function in a two-step procedure. The first step requires, however, using numerical integration as in Amsler et al. (2016).

In the above mentioned papers, there were no environmental variables determining firms' inefficiency. Amsler et al. (2017) provides a systematic treatment of endogeneity in heteroscedastic stochastic frontier models and allows environmental variables to be endogenous because they are correlated with either the statistical noise or the basic inefficiency term or both. When they are only correlated with v_i as in Kutlu (2016), the endogeneity issue is relatively easy to handle. When some environmental variables are endogenous because they are correlated with the random part of u_i neither IV and ML method is simple, because a specific copula must be assumed and simulation methods are necessary to form the IV criterion function or the likelihood.

6.3 The Choice of Orientation: Exogenous and Endogenous Directions

In standard primal and dual representations of firm's technology, researchers often choose between input- and output-oriented measures of firms' inefficiency. However, the distance functions depend on the same vector of inputs and outputs. Both the mathematical programming and econometric regression methods need further qualifications to identify and calculate or estimate the relevant distance functions. In standard DEA, there are models such as the additive formulation that are non-oriented, but traditional oriented models require the specification of different objective functions, either in the envelopment or in the multiplier formulations. In SFA, one needs to select a set of inputs and/or outputs to impose a particular homogeneity or translation property. As Greene (2008, p. 153) notes, the question is which form is appropriate for a given setting? Also, the emergence of new characterizations of the production technology through the directional and GDFs opens a new range of possibilities related to economic behaviour given their flexibility to choose the direction towards the production frontier.

These developments show that the traditional binary choice between input and output orientations is not the only option, unless it is grounded on the firm's economic objective. Indeed, what should lead researchers when deciding on it is the notion of Pareto efficiency and the maximization of utility, which in our current context, and assuming competitive pricing, corresponds either to profit or profitability maximization—ten Raa (2008).

This underlies the choice of duality framework to perform overall economic efficiency analyses (as summarized in Sect. 2.4). For instance, as the input distance function suggests when referring to the degree by which the current input level exceeds the input requirement for production of a particular amount of outputs, it is natural to associate it to (lack of) cost minimization. In this case, it is assumed that inputs are the choice variables and the firm can reduce them at least in the short run without reducing output production. Likewise, as the output distance function suggests when referring to the degree by which output falls short of what can be produced with a given input vector, it is natural to associate this output-oriented function to revenue maximization. In this case, it is assumed that outputs are the choice and adjustable variables. Thus, while the input orientation is intuitive when output is out of control for the firm (e.g. when demand is determined or fixed), the output orientation is intuitive when the inputs are exogenously determined.

Regarding dual representations of firms' technology, the cost approach is preferred if the output side of the firms is exogenous and nondiscretionary, and the opposite is applicable to the revenue side. The choice between profit and profitability (return-to-dollar) is less clear, as both choices are available when both inputs and outputs can be freely adjusted at the discretion of managers. In the short term, managers are normally concerned with attaining maximum profit, but it can be argued that the long-term viability of a firm critically depends on its ability to remain competitive in the market, with profitability representing an economically weighted (by prices) measure of productivity. This is particularly true in markets where the degree of competition is large, firms cannot exert market power, and are expected to make economic profit.

Therefore, the choice of orientation should be determined, at least partially, by the capability of firms to adjust their decisions in order to become fully efficient. However, it should be noted that the distance function concept was originally developed to represent the technology using multiple-input and multiple-output data. Kumbhakar (2012) shows that, while the underlying technology to be modelled is the same, the different orientations only provide different sets of theoretical restrictions to identify the frontier parameters to be estimated. This is also clear in a nonparametric context, where the DEA technology represented by (9) is common to all orientations, while technical efficiency can be measured considering alternative orientations. Moreover, in the SFA framework Kumbhakar, Orea, et al. (2007) show that, once the distance function is known, input (output) oriented

inefficiency scores can be obtained from output (input) distance functions. In a similar manner, Orea et al. (2004) and Parmeter and Kumbhakar (2014; Sect. 4.2) show that both output- and input-oriented inefficiency scores can be computed from an estimated cost function. Thus, if any measure of firms' inefficiency can be estimated using any primal or dual representation of firms' technology, why is the choice of orientation a relevant issue?

It is a relevant issue for at least two empirical reasons. First of all, because both the efficiency scores and the estimated technologies are expected to be different. In the nonparametric DEA framework, Kerstens et al. (2012) and Peyrache and Daraio (2012) study how efficiency results critically depend on the choice of orientation. The latter authors study the sensitivity of the estimated efficiency scores to the directional selection. In the parametric SFA setting, Kumbhakar and Tsionas (2006) and Kumbhakar (2010) also estimate input- and output-oriented stochastic frontier production functions and find that the estimated efficiency, returns to scale, technical change, etc. differ depending on whether one uses the model with input- or output-oriented technical inefficiency. Using a dual approach, Orea et al. (2004) estimate cost frontiers under different specifications which assess how inefficiency enters the data generating process. These authors show that the different models yield very different pictures of the technology and the efficiency levels of the sector, illustrating the importance of choosing the most appropriate model before carrying out production and efficiency analyses. Similar comments can be made if DDFs are used. For instance, Vardanyan and Noh (2006) and Agee et al. (2012) also show that the parameter estimates depend on the choice of the directional vectors.

Second, the choice of orientation is also relevant for the "complexity" of the stochastic part of the model in a SFA model. For instance, Kumbhakar and Tsionas (2006) show that the standard *Maximum Likelihood* (ML) method cannot be applied to estimate input-oriented production functions. They instead use a simulated ML approach as estimation method. Similarly, Orea et al. (2004) estimated stochastic cost frontier models with output-oriented measures of firms' inefficiency using a nonlinear fixed effect approach. If, in contrast, inefficiency is modelled as a one-sided random term, Parmeter and Kumbhakar (2014) show that a stochastic cost frontier model with output-oriented inefficiency is difficult to estimate without additional restrictions on the technology.

In the following subsections, we discuss the choice of orientation from the modelling perspective of the DEA and SFA approaches and summarize the most recent proposals related to the rationale underlying different

possibilities, including those endogenizing the orientation, and driven by the data. This last approach emerges in situations in which there is not an economic or managerial rationale to impose a specific goal.

DEA Framework

As both the directional and GDFs nest the traditional input and output partial orientations, which may not provide the discretion needed in empirical studies that require an independent treatment of both inputs and outputs, we focus our discussion on the DDF (but it can be equivalently extended the GDF).

Clearly, when choosing an orientation, several criteria are on the list. The first one mirrors the rationale behind the input and output distance functions, by setting the orientation for each DMU equal to the observed amounts of inputs and outputs, $\left(-g_x, g_y\right) = (-x, y)$. Färe and Grosskopf (2000, p. 98) justify the choice on the grounds that it provides a link and symmetry with the traditional distance functions as presented above. This implies solving problem (10) substituting the directional vector by the observed amounts, constituting the most common approach in empirical applications relying on the DDF. Alternatively, rather than using individual directions for each firm, it is possible to adopt a so-called egalitarian approach assigning the same direction to all firms. An example of such common direction is to take the average input and output mixes: $\left(-g_x, g_y\right) = (-\bar{x}, \bar{y})$, or the unit vector $\left(-g_x, g_y\right) = (-1, 1)$. Both have the advantage that the direction is neutral. However, the interpretation of the β value corresponding to the distance function is different. When the directional vector is measured in the units of measurement of inputs and outputs, e.g. as with $(-x, y)$ or $(-\bar{x}, \bar{y})$, the efficiency measure corresponds to the *proportion* of the observed input and outputs amounts that is to be detracted and increased to reach the frontier—e.g. for $\beta = 2$, twice the observed amounts, which eases its interpretation. However, its main drawback is that the metric for each observation is different, so their individual values cannot be numerically compared. Also, the DDF is units free in the sense that if we rescale inputs and outputs, as well as their directional vector, by the same vector, then β remains unchanged. However, if the direction is not in the same units of measurement than those of the inputs and outputs, its interpretation differs while the unit free property does not hold. For example, if the unitary vector is chosen, then the distance function β yields the amount in which inputs and outputs need to be decreased and increased to reach the

frontier. This can be easily seen taking the difference between the optimal and observed output and input vectors: i.e. $y + D_T(x, y; -1, 1) \cdot 1_M - y$ and $x + D_T(x, y; -1, 1) \cdot 1_N - x$, and the amount in which outputs can be increased and inputs reduced corresponds to the value of the distance function. This discussion simply shows that since the value of the distance function depends on the directional vector, this should be explicitly considered when choosing a specific metric.

The above directional distance vectors can be considered exogenous, since they are chosen by the researcher based on ad hoc criteria. A second possibility that endogenizes the choice is based on the economic behaviour of the firm. When market prices are observed and firms exhibit an economic optimizing behaviour, Zofío et al. (2013) introduce a profit efficiency measure that projects the evaluated firm to the profit maximizing benchmark, which is in accordance with the ultimate utility-maximizing criteria guiding this chapter, and within the duality framework. Their method searches for a directional vector $(-g_x^*, g_y^*)$ that endogenizes the projection of the evaluated firm so as to achieve that goal—represented by the input-output vector (x^π, y^π) in Fig. 1. The associated profit efficiency measure simultaneously solving the directional vector and identifying the profit-maximizing benchmark given the input and output market prices (w, p) can be calculated in the following way:

$$D_T^*(x_{i'}, y_{i'}; w, p) = \max_{\beta, \lambda_i, g_x, g_y} \beta \tag{36}$$

$$\begin{aligned}
\text{s.t.} \quad & \sum_{i=1}^{I} \lambda_i x_{in} \leq x_{i'n} - \beta g_{x_n}^*, \; n = 1, \ldots, N, \\
& \sum_{i=1}^{I} \lambda_i y_{im} \geq y_{i'm} + \beta g_{y_m}^*, \; m = 1, \ldots, M, \\
& \sum_{m=1}^{M} p_m g_{y_m}^* + \sum_{n=1}^{N} w_n g_{x_n}^* = 1, \\
& \sum_{i=1}^{I} \lambda_i \leq 1, \quad \lambda \in R_+^I.
\end{aligned}$$

By solving this programme, we gain information about firm i's profit inefficiency, the profit maximizing benchmark and the optimal course that it should follow when planning and adopting profit improving strategies.

Programme (36) departs from (10) in two crucial ways. First, as previously remarked, the directional vector is not preassigned and therefore (36) searches for it given the price normalization constraint. Second, the elements of the directional vector (g_x^*, g_y^*) could adopt any value, positive and negative, as long as $(g_x^*, g_y^*) \neq (0_N, 0_M)$. This means that inputs may be increased and outputs reduced when projecting the evaluated firm to the profit maximizing benchmark. The choice of orientation also has relevant consequences when measuring overall economic efficiency according to (7), as the profit normalizing condition is a function of the directional vector $pg_y + wg_x$. Therefore, the proposal by Zofío et al. (2013) normalizing the price constraint to $pg_y^* + wg_x^* = 1$ allows measuring the overall, technical and allocative efficiencies in monetary terms (e.g. dollar valued).

Most importantly, a relevant consequence of this proposal is that it renders the decomposition of overall economic efficiency (7) and (8) redundant when inputs and outputs are fully adjustable at the discretion of managers. This result derives from the fact that the overall economic efficiency is obtained by identifying the profit efficiency measure along the directional vector (g_x^*, g_y^*), which in turn allows determining whether the evaluated firm is on the production frontier or not. Particularly, when $D_T^*(x_{i'}, y_{i'}; w, p) > 0$, so the firm is profit inefficient, and in conjunction with the value of the DDF (10), we can determine whether the source of the inefficiency is technical, $D_T\left(x_{i'}, y_{i'}; g_x, g_y\right) > 0$, or allocative, $D_T\left(x_{i'}, y_{i'}; g_x, g_y\right) = 0$. From a theoretical and conceptual perspective, this proposal solves the *arbitrary* decomposition of profit efficiency as the relative values of the technical and residual allocative efficiencies depend on the exogenous choice of the directional vector. Nevertheless, when some output or inputs are fixed, exogenous or nondiscretionary, it might not be possible to change the production process so as to attain maximum profit, resulting in overall economic decompositions as in (7) and (8).

When selecting a given orientation, several authors, both in the DEA and in the SFA, rely on the existing data to identify the most relevant peers. Based in part on the initial contribution by Payrache and Daraio (2012), Daraio and Simar (2016) proposed a method that allows choosing context specific (or local) directions for firms, considering as benchmarks those facing similar conditions, and without assuming any economic behaviour. These conditions can be associated with the closeness of those benchmark peers to the production (input–output) mix of the evaluated firm or their share of the same contextual conditions (factors), represented by a vector W, e.g. benchmarks facing the same nondiscretionary inputs and

outputs. The method produces an automatic "peer grouping" of the firms as by-products, depending on their comparable circumstances and external conditions. This last feature is what represents the data driven approach these authors refer to.

However, the implementation of the algorithm is complex, as it defines the directional vector in terms of angles in polar coordinates in the multidimensional input–output space, which nevertheless allows these authors to: (1) impose the same direction (angle) using the average of the angles, rather than the average of the observed input and output quantities (so large individual firms will not weigh more in the egalitarian direction) or (2) consider different directions when the contextual factors justify their use. How these external factors in W influence the direction (angles) is carried out through nonparametric regression analysis of the direction on W. It is then a "local" direction determined by its neighbouring (similar) firms. These authors apply their method to simulated and original databases. They compare the results following their data driven method, testing the influence of the external factors included in W and comparing results with those obtained for alternative orientations such as individual-specific distances or the egalitarian approach. The method captures the influence of the contextual factors and provides an efficiency measure that takes into account the particularities of the firms being evaluated with respect to their potential benchmark peers. An implementation of the method in a standard software package would be necessary to popularize this potentially useful, but computationally complex method.

SFA Framework

As mentioned in the introduction of this section, the choice of orientation in SFA shows its relevance with respect to the "complexity" of the stochastic part of the model. To see this clearly, assume that we want to estimate firms' technology using a stochastic input-oriented distance function $D_I\left(x_i^*, y_i^*\right)e^{v_i} = 1$, where the asterisk stands for efficient units and v_i is the traditional noise term. If inefficient production is measured in terms of output reductions (in this case, we assume that $x_i = x_i^*$), the model to be estimated can be written after imposing linear homogeneity in inputs as:

$$-\ln x_{1i} = \ln D_I\left(\frac{x_i}{x_{1i}}, y_1 e^{u_i}\right) + v_i \tag{37}$$

As customary, if we assume that the distance function has a flexible functional form such as the Translog and the firm only produces a single output, the model to be estimated can be written as:

$$\begin{aligned} -\ln x_{1i} = \beta_0 + \sum_{n=2}^{N} \beta_n \ln\left(x_{ni}/x_{1i}\right) + \frac{1}{2}\sum_{n=2}^{N}\sum_{n'=2}^{N} \beta_{nn'} \ln\left(x_{ni}/x_{1i}\right) \ln\left(x_{n'i}/x_{1i}\right) \\ + \beta_y \ln y_i + \frac{1}{2}\beta_{yy} \ln y_i^2 + \sum_{n=2}^{N} \beta_{ny} \ln\left(x_{ni}/x_{1i}\right) \ln y_i + \varepsilon_i, \end{aligned} \tag{38}$$

where

$$\varepsilon_i = v_i + \left[\beta_y + \sum_{n=2}^{N} \beta_{ny} \ln\left(x_{ni}/x_{1i}\right)\right] u_i + \frac{1}{2}\beta_{yy} u_i^2. \tag{39}$$

The composed error term ε_i involves three random terms, v_i, u_i and u_i^2. As Parmeter and Kumbhakar (2014) point out in a cost setting, the presence of the u_i^2 term in ε_i makes the derivation of a closed likelihood function impossible. Thus, this precludes using standard maximum likelihood techniques to obtain the parameter estimates. Similar comments can be made if we were to use a directional or generalized−hyperbolic distance function. In all cases where we have intractable likelihood functions, they can be maximized by simulated maximum likelihood.[40] A final important remark regarding Eqs. (38) and (39) is that the input orientation of the distance function does not force the researcher to use an input-oriented measure of firms' inefficiency. We first do it just for simplicity and in doing so are likely to attenuate endogeneity problems as well. The same remark obviously can be made for other primal (and dual) representations of firms' technology.

So far, we have implicitly assumed that the researcher selects a particular orientation before carrying out production and efficiency analyses. The selection is normally based on the features of the industry being examined, e.g. on whether input or outputs are exogenously determined. However, as in its nonparametric DEA counterpart, the input–output orientation issue may also be viewed as a data driven issue, and thus, the decision can be based on performing proper model selection tests. If we allow for more than one orientation, these tests select the orientation that better fit the data or the orientation that provides the largest value for the likelihood function

[40]As shown by Parmeter and Kumbhakar (2014, p. 52) using a Translog cost function, if the production technology is homogeneous in outputs, the model can be estimated using simple ML techniques.

given the number of parameters to be estimated. For instance, Orea et al. (2004) fit input-, output- and hyperbolic-oriented cost frontier models using a panel data set on Spanish dairy farms. The performed tests show that the input-oriented model is the best among the models estimated, a result that is consistent with the fact that this provides the most credible estimates of scale economies given the structure of the sector.

In the SFA framework, traditional output- and input-oriented models impose a common orientation for all firms over time. The same happens in the paper mentioned above. Kumbhakar et al. (2007) note that this could be a strong assumption in some applications. That is, a given firm could be operating in either regime at any time. These authors treat the input and output distance functions as two latent regimes in a finite mixture model, representing firms' technology by a general stochastic distance function: $0 = \ln D(x_i, y_i, \beta) + v_i + su_i$, where β is a vector of technological parameters, and u_i is a one-sided random variable representing technical inefficiency whose sign (i.e. $s = 1$ or $s = -1$) depends on the chosen orientation. The determination of the efficiency orientation for each firm is addressed by adopting a latent class structure so that the technologies and the probability of being in the input-/output-oriented inefficiency model are estimated simultaneously by ML. The contribution of firm i to the likelihood is:

$$LF_i = LF_i^I(\gamma_I)\Pi_i(\theta) + LF_i^O(\gamma_O)(1 - \Pi_i(\theta)), \tag{40}$$

where $\gamma = (\beta, \sigma_v, \sigma_u)$ is the whole set of parameters of the stochastic distance function, $LF_i^I(\gamma_I)$ is the likelihood function of an input distance function model, $LF_i^O(\gamma_O)$ is the likelihood function of an output distance function model, $\Pi_i(\theta)$ is the probability of being in the input-oriented class, and $1 - \Pi_i(\theta)$ is the probability of being in the output-oriented class. The computed posterior probabilities are then used to know whether a particular firm is maximizing output (revenue) or minimizing input use (cost). In essence, the Kumbhakar et al. (2007) model allows the data to sort themselves into the input- and output-oriented regimes rather than arbitrarily assuming that all observations obey one or the other at the outset.

The latent class model used by these authors allows different orientations in an exogenous fashion. There are more probable efficiency measures than others, but the latent class structure of the model does not allow firms to choose the orientation (i.e. the economic objective) they wish to pursue. Therefore, one interesting extension of this model is to endogenize the selection of the orientation of the efficiency measures. This likely can be carried out by adapting one of the models recently introduced in the SFA literature

to deal with sample selection problems for this setting (see, e.g., Greene 2010, Kumbhakar et al. 2009, and Lai 2013). The key feature of these models is that production technology is a decision made by the firm itself and thus renders the sample split variable endogenous. The direct consequence of ignoring the endogeneity of the sample split variable is the estimation bias of the production technology, even if the differences in technology (in our case, efficiency orientations) are allowed in the model.

Atkinson and Tsionas (2016, 2018) pursue a similar objective using DDFs. The typical fixed direction approach often assumes +1 directions for outputs and −1 directions for inputs. They argue, however, that since goods (inputs) are produced (demanded) by firms, their relative valuation may not be 1-to-1 for all firms. They generalize the standard (and restricted) models by jointly estimating a quadratic technology-oriented DDF, not with directions chosen a priori, but with chosen optimal directions that are consistent with cost minimization or profit maximization. In particular, providing the parametric SFA counterpart to Zofío et al. (2013), they first consider the typically employed Quadratic directional distance function of all inputs and outputs, and next, they append price equations to obtain a cost-minimization or profit-maximization DDF system. Therefore, they also generalize the dual relationship between the profit function and the technology oriented DDF, as established by Chambers (1998). These equations allow identification of directions for each input and output. They estimate their DDF systems using respectively Bayesian and GMM techniques, obtaining estimates of all structural parameters *and* optimal directions.

7 Concluding Remarks

This contribution serves as guide to efficiency evaluation from an economic perspective and, complementing several chapters in this handbook, intends to make the reader aware of the different alternatives available for choice when undertaking research in the field. The analytical framework relies on the most general models and up to date representations of the production technology and economic performance through directional and GDFs, nesting the traditional approaches well known in the literature, while complementing them with current issues related to their empirical implementation.

In this chapter, we stress the importance of choosing a suitable analytical framework that is in accordance with the industry characteristics and the

restrictions faced by the firm, most particularly the relative discretion that managers have over output production and input usage. This sets the stage for the economic objective of the firm that, in an unconstrained setting, is assumed to maximize profit or profitability, both of which can be related to cost minimization and revenue maximization. Therefore, the underlying principle in the measurement of economic efficiency and the necessary choice of orientation for flexible distance functions is that of Pareto efficiency, i.e. utility maximization, which indeed corresponds to the above objectives under the assumption of competitive market prices. Once the theoretical foundation for the measurement of overall economic efficiency is determined, the next question that scholars face is the choice of methods that are available to study variability in firm performance. Following previous chapters, we discuss the main characteristics, pros and cons and relevant assumptions that need to be made to successfully undertake a study using either DEA or SFA techniques. As all concerns discussed here are shared by both approaches, we do not add to the almost endless debate on which approach is best, loosely based on their relative strengths and weaknesses, but advice the reader on the capabilities of each method to better address the existing empirical limitations and deal with research constraints.

We conclude emphasizing the relevance of the methods surveyed in this chapter in unveiling the economic performance of firm in terms of technical and allocative (in)efficiencies, whose persistence and variability call for further integration within the discipline of industrial organization. Efficiency and productivity analysis is now part of the toolbox in regulation and competition theory, providing the necessary analytical and quantitative results that allow the setting of firms' incentives in regulated industries (Agrell and Bogetoft 2013), the evaluation of firms' market power through mark-ups (Abhiman and Kumbhakar 2016; Orea and Steinbuks 2018) or the effects of mergers and acquisitions from the perspective of competition theory (Fiordelisi 2009). Nevertheless, it is possible to think of additional fields where firms' heterogeneity in terms of their relative productivity is fundamental, as in the new trade models proposed by Melitz and Ottaviano (2008), where trade openness among countries triggers the Darwinian process of firm selection in domestic markets, with those situated in the lower tail of the (in)efficiency distribution exiting the industry. It is by now clear that the homogeneity associated with the canonical model of perfect competition is giving way to the reality associated with the indisputable evidence of inefficient behaviour. On these grounds, in terms of economic, technical and allocative fundamentals, the pieces of the inefficiency puzzle go towards explaining why firms deviate from best practice operations and, in

this sense, make a valuable contribution to a wide range of research issues. As shown in this handbook, many challenges are still ahead, but cross-fertilization of ideas with other research fields will result in a better understanding of the ultimate causes and consequences of inefficient economic performance.

References

Abhiman, D., and S.C. Kumbhakar. 2016. Markup and efficiency of Indian Banks: An input distance function approach. *Empirical Economics* 51 (4): 1689–1719.

Aczél, J. 1966. *Lectures on functional equations and their applications*. New York: Academic Press.

Adler, N., and B. Golany. 2002. Including principal component weights to improve discrimination in data envelopment analysis. *Journal of the Operational Research Society* 53: 985–991.

Adler, N., and E. Yazhemsky. 2010. Improving discrimination in data envelopment analysis: PCA-DEA or variable reduction. *European Journal of Operational Research* 202: 273–284.

Adragni, K.P., and D. Cook. 2009. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society* 367: 4385–4405.

Agee, M.D., S.E. Atkinson, and T.D. Crocker. 2012. Child maturation, time-invariant, and time-varying inputs: Their interaction in the production of child human capital. *Journal of Productivity Analysis* 35: 29–44.

Agrell, P.J., and P. Bogetoft. 2013. Benchmarking and regulation. Discussion Paper No. 8, Center for Operations Research and Econometrics, Université Catholique de Louvain, Belgium.

Aigner, D.J., C.A. Lovell, and P. Schmidt. 1977. Formulation and estimation of stochastic frontier production functions. *Journal of Econometrics* 6 (1): 21–37.

Almanidis, P., J. Qian, and R.C. Sickles. 2010. Bounded stochastic frontiers with an application to the US banking industry: 1984–2009. Unpublished Manuscript, Rice University. http://economics.rice.edu/WorkArea/DownloadAsset.aspx?id=-497.

Álvarez, I., J. Barbero, and J.L. Zofio. 2016. A data envelopment analysis toolbox for MATLAB. Working Papers in Economic Theory 2016/03, Department of Economics, Universidad Autónoma de Madrid, Spain. www.deatoolbox.com. Forthcoming in the *Journal of Statistical Software*.

Álvarez, A., C. Amsler, L. Orea, and P. Schmidt. 2006. Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics. *Journal of Productivity Analysis* 25: 201–212.

Amsler, C., A. Prokhorov, and P. Schmidt. 2014. Using copulas to model time dependence in stochastic frontier models. *Econometric Reviews* 33 (5–6): 497–522.

Amsler, C., A. Prokhorov, and P. Schmidt. 2016. Endogeneity in stochastic frontier models. *Journal of Econometrics* 190 (2): 280–288.

Amsler, C., A. Prokhorov, and P. Schmidt. 2017. Endogenous environmental variables in stochastic frontier models. *Journal of Econometrics* 199: 131–140.

Aparicio, J., and J.L. Zofío. 2017. Revisiting the decomposition of cost efficiency for non-homothetic technologies: A directional distance function approach. *Journal of Productivity Analysis* 48 (2–3): 133–146.

Aparicio, J., J.T. Pastor, and J.L. Zofío. 2015. How to properly decompose economic efficiency using technical and allocative criteria with non-homothetic DEA technologies. *European Journal of Operational Research* 240 (3): 882–891.

Aparicio, J., J.T. Pastor, and J.L. Zofío. 2017. Can Farrell's allocative efficiency be generalized by the directional distance function approach? *European Journal of Operational Research* 257 (1): 345–351.

Aparicio, J., F. Borras, J.T. Pastor, and J.L. Zofío. 2016. Loss distance functions and profit function: General duality results. In *Advances in efficiency and productivity*, ed. J. Aparicio, C.A.K. Lovell, and J.T. Pastor, 71–98. New York: Springer.

Atkinson, S.E., and M.G. Tsionas. 2016. Directional distance functions: Optimal endogenous directions. *Journal of Econometrics* 190 (2): 301–314.

Atkinson, S.E., and M.G. Tsionas. 2018. Shadow directional distance functions with bads: GMM estimation of optimal directions and efficiencies. *Empirical Economics* 54 (1): 207–230.

Balk, B.M. 1998. *Industrial price, quantity, and productivity indices: The micro-economic theory and an application.* Dordrecht: Kluwer Academic Publishers.

Bandyopadhyay, D., and A. Das. 2006. On measures of technical inefficiency and production uncertainty in stochastic frontier production model with correlated error components. *Journal of Productivity Analysis* 26: 165–180.

Banker, R.D., and R. Morey. 1986. Efficiency analysis for exogenously fixed inputs and outputs. *Operation Research* 34 (4): 513–521.

Banker, R.D., and R. Natarajan. 2008. Evaluating contextual variables affecting productivity using data envelopment analysis. *Operations Research* 56 (1): 48–58.

Banker, R.D., and R.M. Thrall. 1992. Estimation of returns to scale using data envelopment analysis. *European Journal of Operational Research* 62 (1): 74–84.

Banker, R.D., A. Charnes, W.W. Cooper, J. Swarts, and D. Thomas. 1989. An introduction to data envelopment analysis with some of its models and their uses. *Research in Government and Nonprofit Accounting* 5: 125–163.

Bauer, P.W., A.N. Berger, G.D. Ferrier, and D.B. Humphrey. 1998. Consistency conditions for regulatory analysis of financial institutions: A comparison of frontier efficiency methods. *Journal of Economics and Business* 50 (2): 85–114.

Bravo-Ureta, B., D. Solís, V. Moreira-López, J. Maripani, A. Thiam, and T. Rivas. 2007. Technical efficiency in farming: A meta-regression analysis. *Journal of Productivity Analysis* 27 (1): 57–72.

Brons, M., P. Nijkamp, E. Pels, and P. Rietveld. 2005. Efficiency of urban public transit: A meta analysis. *Transportation* 32 (1): 1–21.

Bura, E. 2003. Using linear smoothers to assess the structural dimension of regressions. *Statistica Sinica* 13: 143–162.

Bura, E., and J. Yang. 2011. Dimension estimation in sufficient dimension reduction: A unifying approach. *Journal of Multivariate Analysis* 102: 130–142.

Carta, A., and M.F.J. Steel. 2012. Modelling multi-output stochastic frontiers using copulas. *Computational Statistics & Data Analysis* 56 (11): 3757–3773.

Chambers, R.G. 1988. *Applied production analysis: A dual approach*. New York: Cambridge University Press.

Chambers, R.G. 1998. Input and output indicators. In *Index numbers in honour of Sten Malmquist*, ed. R. Färe, S. Grosskopf, and R.R. Russell, 241–272. Boston: Kluwer Academic Publishers.

Chambers, R.G., and J. Quiggin. 2000. *Uncertainty, production, choice and agency: The state-contingent approach*. New York: Cambridge University Press.

Chambers, R.G., Y. Chung, and R. Färe. 1996. Benefit and distance functions. *Journal of Economic Theory* 70: 407–419.

Chambers, R.G., Y. Chung, and R. Färe. 1998. Profit, directional distance functions and Nerlovian efficiency. *Journal of Optimization Theory and Applications* 95 (2): 351–364.

Chavas, J.P., and T.M. Cox. 1999. A generalized distance function and the analysis of production efficiency. *Southern Economic Journal* 66 (2): 295–318.

Coelli, T., and S. Perelman. 1996. Efficiency measurement, multiple-output technologies and distance functions: With application to European railways. No. DP 1996/05. CREPP.

Cook, R.D., and L. Ni. 2005. Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association* 100: 410–428.

Cook, W.D., and J. Zhu. 2015. DEA Cross Efficiency. In *Data envelopment analysis: A handbook of models and methods*, ed. J. Zhu. New York: Springer.

Cooper, W.W., L.M. Seiford, and K. Tone. 2007. *Data envelopment analysis: A comprehensive text with models, applications, references and DEA-Solver software.* Princeton, NJ: Princeton University Press.

Cordero, J.M., D. Santín, and G. Sicilia. 2015. Testing the accuracy of DEA estimates under endogeneity through a Monte Carlo simulation. *European Journal of Operational Research* 244 (2): 511–518.

Cuesta, R., and J.L. Zofío. 2005. Hyperbolic efficiency and parametric distance functions: With application to Spanish savings banks. *Journal of Productivity Analysis* 24 (1): 31–48.

Cummins, J.D., and H. Zi. 1998. Comparison of frontier efficiency methods: An application to the U.S. life insurance industry. *Journal of Productivity Analysis* 10: 131–152.

Daraio, C., and L. Simar. 2005. Introducing environmental variables in nonparametric frontier models: A probabilistic approach. *Journal of Productivity Analysis* 24 (1): 93–121.

Daraio, C., and L. Simar. 2016. Efficiency and benchmarking with directional distances: A data-driven approach. *Journal of the Operational Research Society* 67 (7): 928–944.

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

Diewert, W.E. 1971. An application of the Shephard duality theorem: A generalized Leontief production function. *Journal of Political Economy* 79: 461–507.

Dyson, R.G., R. Allen, A.S. Camanho, V.V. Podinovski, C.S. Sarrico, and E.A. Shale. 2001. Pitfalls and protocols in DEA. *European Journal of Operational Research* 132 (2): 245–259.

Emvalomatis, G. 2009. Parametric models for dynamic efficiency measurement. Unpublished thesis.

Fan, J., and J. Lv. 2010. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20 (1): 101–148.

Färe, R., and D. Primont. 1995. *Multi-output production and duality: Theory and applications*. Boston: Kluwer Academic Publishers.

Färe, R., and S. Grosskopf. 2000. Notes on some inequalities in economics. *Economic Theory* 15 (1): 227–233.

Färe, R., S. Grosskopf, and C.A.K. Lovell. 1985. *The measurement of efficiency of production*. Boston, USA: Kluwer-Nijhoff.

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

Färe, R., S. Grosskopf, D.W. Noh, and W. Weber. 2005. Characteristics of a polluting technology: Theory and practice. *Journal of Econometrics* 126: 469–492.

Farrell, M. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A, General* 120 (3): 253–281.

Filippini, M., and W. Greene. 2016. Persistent and transient productive inefficiency: A maximum simulated likelihood approach. *Journal of Productivity Analysis* 45 (2): 187–196.

Fiordelisi, F. 2009. *Mergers and acquisitions in European banking*. New York: Palgrave Macmillan.

Fisher, R.A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society* 222: 309–368.

Fonseca, J.R.S., and M.G.M.S. Cardoso. 2007. Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis* 11 (2): 155–173.

Fried, H.O., C.A.K. Lovell, and S.S. Shelton. 2008. *The measurement of productive efficiency and productivity growth*. New York: Oxford University Press.

Fried, H.O., C.A.K. Lovell, S.S. Schmidt, and S. Yaisawarng. 2002. Accounting for environmental effects and statistical noise in data envelopment analysis. *Journal of Productivity Analysis* 17: 157–174.

Friedman, L., and Z. Sinuany-Stern. 1997. Scaling units via the canonical correlation analysis in the DEA context. *European Journal of Operational Research* 100 (3): 25–43.

Golany, B., and Y. Roll. 1989. An application procedure for DEA. *Omega* 17 (3): 237–250.

Greene, W. 2005. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics* 126: 269–303.

Greene, W. 2008. The econometric approach to efficiency analysis. In *The measurement of productive efficiency and productivity growth*, ed. H. Fried, C.A. Lovell, and S.S. Schmidt, 92–250. New York: Oxford University Press.

Greene, W.H. 1990. A gamma-distributed stochastic frontier model. *Journal of Econometrics* 46 (1–2): 141–164.

Greene, W.H. 2010. A stochastic frontier model with correction for sample selection. *Journal of Productivity Analysis* 34 (1): 15–24.

Griffiths, W.E., and G. Hajargasht. 2016. Some models for stochastic frontiers with endogeneity. *Journal of Econometrics* 190 (2): 341–348.

Griffiths, W.E., C.J. O'Donnell, and A. Tan-Cruz. 2000. Imposing regularity conditions on a system of cost and factor share equations. *Australian Journal of Agricultural and Resource Economics* 44 (1): 107–127.

Growitsch, C., T. Jamasb, and H. Wetzel. 2012. Efficiency effects of observed and unobserved heterogeneity: Evidence from Norwegian electricity distribution networks. *Energy Economics* 34 (2): 542–548.

Guan, Z., S.C. Kumbhakar, R.J. Myers, and A.O. Lansink. 2009. Measuring excess capital capacity in agricultural production. *American Journal of Agricultural Economics* 91: 765–776.

Hjalmarsson, L., S.C. Kumbhakar, and A. Heshmati. 1996. DEA, DFA and SFA: A Comparison. *Journal of Productivity Analysis* 7 (2): 303–327.

Huang, T.H., D.L. Chiang, and S.W. Chao. 2017. A new approach to jointly estimating the lerner index and cost efficiency for multi-output banks under a stochastic meta-frontier framework. *Quarterly Review of Economics and Finance* 65: 212–226.

Jenkins, L., and M. Anderson. 2003. A multivariate statistical approach to reducing the number of variables in data envelopment analysis. *European Journal of Operational Research* 147: 51–61.

Johnson, A.L., and T. Kuosmanen. 2012. One-stage and two-stage DEA estimation of the effects of contextual variables. *European Journal of Operational Research* 220 (2): 559–570.

Johnson, A.L., and T. Kuosmanen. 2015. An introduction to CNLS and StoNED methods for efficiency analysis: Economic insights and computational aspects. In *Benchmarking for performance evaluation: A production frontier approach*, ed. S.C. Ray, S.C. Kumbhakar, and P. Dua. New Delhi: Springer.

Jondrow, J., C.A. Lovell, S. Materov, and P. Schmidt. 1982. On the estimation of technical efficiency in the stochastic frontier production function model. *Journal of Econometrics* 19 (2–3): 233–238.

Kerstens, K., A. Mounir, and I. Van de Woestyne. 2012. Benchmarking Mean-Variance Portfolios using a shortage function: The choice of direction vector affects rankings! *Journal of the Operational Research Society* 63 (9): 1199–1212.

Kim, M. 1986. Banking technology and the existence of a consistent output aggregate. *Journal of Monetary Economics* 18 (2): 181–195.

Kittelsen, S.A.C. 1993. Stepwise DEA: Choosing variables for measuring technical efficiency in Norwegian electricity distribution. Memorandum No. 06/93, Department of Economics, University of Oslo, Norway.

Koopmans, T. 1951. An analysis of production as an efficient combination of activities. In *Activity analysis of production and allocation*. Cowles Commission for Research in Economics, Monograph. 13, ed. T. Koopmans. New York: Wiley.

Kumbhakar, S. 2012. Specification and estimation of primal production models. *European Journal of Operational Research* 217 (4): 509–518.

Kumbhakar, S.C. 2010. Efficiency and productivity of world health systems: Where does your country stand? *Applied Economics* 42 (13): 1641–1659.

Kumbhakar, S.C. 2011. Estimation of production technology when the objective is to maximize return to the outlay. *European Journal of Operations Research* 208 (2): 170–176.

Kumbhakar, S.C., and C.A. Lovell. 2000. *Stochastic frontier analysis*. Cambridge: Cambridge University Press.

Kumbhakar, S.C., and E.G. Tsionas. 2006. Estimation of stochastic frontier production functions with input-oriented technical efficiency. *Journal of Econometrics* 133 (1): 71–96.

Kumbhakar, S.C., and E.G. Tsionas. 2011. Stochastic error specification in primal and dual production systems. *Journal of Applied Econometrics* 26: 270–297.

Kumbhakar, S.C., E.G. Tsionas, and T. Sipiläinen. 2009. Joint estimation of technology choice and technical efficiency: An application to organic and conventional dairy farming. *Journal of Productivity Analysis* 31 (2): 151–161.

Kumbhakar, S.C., F. Asche, and R. Tveteras. 2013. Estimation and decomposition of inefficiency when producers maximize return to the outlay: An application to Norwegian fishing trawlers. *Journal of Productivity Analysis* 40: 307–321.

Kumbhakar, S.C., W. Hung-Jen, and A.P. Horncastle. 2015. *A Practitioner's guide to stochastic frontier analysis using Stata*. Cambridge: Cambridge University Press.

Kumbhakar, S.C., B.U. Park, L. Simar, and E.G. Tsionas. 2007. Nonparametric stochastic frontiers: A local maximum likelihood approach. *Journal of Economics* 137 (1): 1–27.

Kumbhakar, S.C., L. Orea, A. Rodríguez-Álvarez, and E.G. Tsionas. 2007. Do we have to estimate an input or an output distance function? An application of the mixture approach to European railways. *Journal of Productivity Analysis* 27 (2): 87–100.

Kutlu, L. 2010. Battese-Coelli estimator with endogenous regressors. *Economic Letters* 109: 79–81.

Kutlu, L. 2016. A time-varying true individual effects model with endogenous regressors. Unpublished manuscript.

Lai, H.P. 2013. Estimation of the threshold stochastic frontier model in the presence of an endogenous sample split variable. *Journal of Productivity Analysis* 40 (2): 227–237.

Lai, H.P., and C.J. Huang. 2010. Likelihood ratio tests for model selection of stochastic frontier models. *Journal of Productivity Analysis* 34 (1): 3–13.

Lau, L.J. 1986. Functional forms in econometric model building. *Handbook of Econometrics* 3: 1515–1566.

Lewin, A.Y., R.C. Morey, and T.J. Cook. 1982. Evaluating the administrative efficiency of courts. *Omega* 10 (4): 401–411.

Li, K. 1991. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* 86: 316–342.

Li, Q. 1996. Estimating a stochastic production frontier when the adjusted error is symmetric. *Economics Letters* 52: 221–228.

Li, Q., C. Huang, D. Li, and T. Fu. 2002. Semiparametric smooth coefficient models. *Journal of Business and Economic Statistics* 20 (3): 412–422.

Luenberger, D.G. 1992. New optimality principles for economic efficiency and equilibrium. *Journal of Optimization Theory and Applications* 75 (2): 221–264.

Malikov, E., S.C. Kumbhakar, and M.G. Tsionas. 2015. A cost system approach to the stochastic directional technology distance function with undesirable outputs: The case of US banks in 2001–2010. *Journal of Applied Econometrics* 31 (7): 1407–1429.

Meeusen, W., and J. van den Broeck. 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* 18 (2): 435–444.

Melitz, M.J., and G.I.P. Ottaviano. 2008. Market size, trade and productivity. *Review of Economic Studies* 75 (1): 295–2316.

Minkowski, H. 1911 *Theorie der Konvexen Körper*. Gesammelte Abhandlungen II. Leipzig and Berlin: B.G. Teubner.

Mittelhammer, R.C., G.G. Judge, and D.J. Miller. 2000. *Econometric foundations*. Cambridge: Cambridge University Press.

Naik, P.A., M.R. Hagerty, and C.L. Tsai. 2000. A new dimension reduction approach for data-rich marketing environments: Sliced inverse regression. *Journal of Marketing Research* 37 (1): 88–101.

Nerlove, M. 1965. *Estimation and identification of Cobb-Douglas production functions*. Chicago: Rand McNally & Co.

Nieswand, M., A. Cullmann, and A. Neumann. 2009. Overcoming data limitations in nonparametric benchmarking: Applying PCA-DEA to natural gas transmission. DIW Discussion Papers, No. 962.

Norman, M., and B. Stoker. 1991. *Data envelopment analysis: The assessment of performance*. New York: Wiley.

Nunamaker, T.R. 1985. Using data envelopment analysis to measure the efficiency of non-profit organizations: A critical evaluation. *Managerial and Decision Economics* 6 (1): 50–58.

Odeck, J., and S. Brathen. 2012. A meta-analysis of DEA and SFA studies of the technical efficiency of seaports: A comparison of fixed and random-effects regression models. *Transportation Research Part A: Policy and Practice* 46 (10): 1574–1585.

O'Donnell, C.J., and T.J. Coelli. 2005. A Bayesian approach to imposing curvature on distance functions. *Journal of Econometrics* 126 (2): 493–523.

O'Donnell, C.J., R.G. Chambers, and J. Quiggin. 2010. Efficiency analysis in the presence of uncertainty. *Journal of Productivity Analysis* 33 (1): 1–17.

Orea, L. 2002. A parametric decomposition of a generalized Malmquist productivity index. *Journal of Productivity Analysis* 18: 5–22.

Orea, L., and I. Álvarez. 2019. A new stochastic frontier model with crosssectional effects in both noise and inefficiency terms. *Journal of Econometrics* (Forthcoming). https://doi.org/10.1016/j.jeconom.2019.07.004.

Orea, L., and S. Kumbhakar. 2004. Efficiency measurement using stochastic frontier latent class model. *Empirical Economics* 29 (1): 169–183.

Orea, L., and J. Steinbuks. 2018. Estimating market power in homogenous product markets using a composed error model: Application to the california electricity market. *Economic Inquiry*. https://doi.org/10.1111/ecin.12539.

Orea, L., and J.L. Zofío. 2017. A primer on the theory and practice of efficiency and productivity analysis. Efficiency Series Papers 2017/05, Department of Economics, Oviedo Efficiency Group (OEG), University of Oviedo, Spain.

Orea, L., C. Growitsch, and J. Jamasb. 2015. Using supervised environmental composites in production and efficiency analyses: An application to Norwegian electricity networks. *Competition and Regulation in Network Industries* 16 (3): 260–288.

Orea, L., M. Llorca, and M. Filippini. 2015. A new approach to measuring the rebound effect associated to energy efficiency improvements: An application to the US residential energy demand. *Energy Economics* 49: 599–609.

Orea, L., D. Roibás, and A. Wall. 2004. Choosing the technical efficiency orientation to analyze firms technology: A model selection test approach. *Journal of Productivity Analysis* 22 (1–2): 51–71.

Orme, C., and P. Smith. 1996. The potential for endogeneity bias in data envelopment analysis. *Journal of the Operational Research Society* 47 (1): 73–83.

Parmeter, C.F., and S.C. Kumbhakar. 2014. Efficiency analysis: A primer on recent advances. *Foundations and Trends in Econometrics* 7 (3–4): 191–385.

Pastor, J.T., J.L. Ruiz, and I. Sirvent. 2002. A statistical test for nested radial DEA models. *Operations Research* 50 (4): 728–735.

Peyrache, A., and T. Coelli. 2009. Testing procedures for detection of linear dependencies in efficiency models. *European Journal of Operational Research* 198 (2): 647–654.

Peyrache, A., and C. Daraio. 2012. Empirical tools to assess the sensitivity of directional distance functions to direction selection. *Applied Economics* 44 (8): 933–943.

Podinovski, V.V. 2015. DEA models with production trade-offs and weight restrictions. In *Data envelopment analysis: A handbook of models and methods*, ed. J. Zhu, 105–144. New York: Springer.

Ray, S.C. 1988. Data envelopment analysis, nondiscretionary inputs and efficiency: An alternative interpretation. *Socio-Economic Planning Science* 22 (4): 167–176.

Ruggiero, J. 1996. On the measurement of technical efficiency in the public sector. *European Journal of Operational Research* 90: 553–565.

Santín, D., and G. Sicilia. 2017. Dealing with endogeneity in data envelopment analysis applications. *Expert Systems with Applications* 68: 173–184.

Sengupta, J.K. 1990. Tests of efficiency in data envelopment analysis. *Computers & Operations Research* 17 (2): 123–132.

Sexton, T.R., R.H. Silkman, and A.J. Hogan. 1986. Data envelopment analysis: Critique and extension. In *New directions for program evaluation*, ed. R.H. Silkman, 73–105. San Francisco: Jossey-Bass.

Shephard, R.W. 1970. *Theory of cost and production functions*. Princeton, NJ: Princeton University Press.

Simar, L., and P.W. Wilson. 2007. Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics* 136 (1): 31–64.

Simar, L., and P.W. Wilson. 2010. Inferences from cross-sectional, stochastic frontier models. *Econometric Reviews* 29: 62–98.

Sinuany-Stern, Z., and L. Friedman. 1998. DEA and the discriminant analysis of ratios for ranking units. *European Journal of Operational Research* 111 (3): 470–478.

Smith, M.D. 2008. Stochastic frontier models with dependent error components. *Econometrics Journal* 11: 172–192.

Stevenson, R.E. 1980. Likelihood functions for generalized stochastic frontier estimation. *Journal of Econometrics* 13 (1): 57–66.

Sun, K., and S.C. Kumbhakar. 2013. Semiparametric Smooth-Coefficient stochastic frontier model. *Economics Letters* 120: 305–309.

ten Raa, T. 2008. Debreu's coefficient of resource utilization, the solow residual, and TFP: The connection by Leontief preferences. *Journal of Productivity Analysis* 30: 191–199.

ten Raa, T. 2011. Benchmarking and industry performance. *Journal of Productivity Analysis* 36: 258–292.

Thompson, R.G., F.D. Singleton, R.M. Thrall, and B.A. Smith. 1986. Comparative site evaluations for locating a high-energy physics lab in Texas. *Interfaces* 16: 35–49.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B* 58 (1): 267–288.

Tran, K.C., and E.G. Tsionas. 2015. Endogeneity in stochastic frontier models: Copula approach without external instruments. *Economics Letters* 133: 85–88.

Tsionas, E.G., S.C. Kumbhakar, and E. Malikov. 2015. Estimation of input distance functions: A system approach. *American Journal of Agricultural Economics* 97 (5): 1478–1493.

Ueda, T., and Y. Hoshiai. 1997. Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs. *Journal of the Operational Research Society of Japan* 40: 466–478.

Vardanyan, M., and D.W. Noh. 2006. Approximating pollution abatement costs via alternative specifications of a multi-output production technology: A case of the U.S. electric utility industry. *Journal of Environmental Management* 80 (2): 177–190.

Wagner, J.M., and D.G. Shimshak. 2007. Stepwise selection of variables in data envelopment analysis: Procedures and managerial perspectives. *European Journal of Operational Research* 180: 57–67.

Wang, H.J. 2002. Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model. *Journal of Productivity Analysis* 18 (3): 241–253.

Wang, H.J., and C.W. Ho. 2010. Estimating fixed-effect panel stochastic frontier models by model transformation. *Journal of Econometrics* 157 (2): 286–296.

Wang, H.J., and P. Schmidt. 2002. One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis* 18: 129–144.

Wang, W.S., and P. Schmidt. 2009. On the distribution of estimated technical efficiency in stochastic frontier models. *Journal of Econometrics* 148: 36–45.

Wilson, P.W. 2003. Testing independence in models of productive efficiency. *Journal of Productivity Analysis* 20 (3): 361–390.

Xia, Y., H. Tong, W.K. Li, and L.X. Zhu. 2002. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B* 64: 363–410.

Yu, W., T. Jamasb, and M. Pollitt. 2009. Does weather explain cost and quality performance? An analysis of UK electricity distribution companies. *Energy Policy* 37 (11): 4177–4188.

Zhu, J. 1998. Data envelopment analysis vs. principle component analysis: An illustrative study of economic performance of Chinese cities. *European Journal of Operational Research* 11: 50–61.

Zofío, J.L., J. Pastor, and J. Aparicio. 2013. The directional profit efficiency measure: On why profit inefficiency is either technical or allocative. *Journal of Productivity Analysis* 40 (3): 257–266.

Zofío, J.L., and A.M. Prieto. 2006. Return to dollar, generalized distance function and the fisher productivity index. *Spanish Economic Review* 8 (2): 113–138.

Pricing Inputs and Outputs: Market Prices Versus Shadow Prices, Market Power, and Welfare Analysis

Aditi Bhattacharyya, Levent Kutlu and Robin C. Sickles

1 Introduction

Giving policy advices solely based on market prices may be misleading when the prices give distorted signals, i.e., diverge from socially efficient prices. Potential reasons for market prices to diverge from efficiency prices include, but not limited to, controlled prices, externalities, imperfect competition, taxes, trade controls, etc. Broadly, market failure may occur due to structure (characteristics) of the market or government intervention.

Imperfect competition (either in input or in output markets) or externalities are causes of market failures due to the structure of the market. Most markets face some forms of imperfect competition. For example, in a market with tacitly colluding firms or a natural monopoly, the prices deviate from the socially optimal prices. An externality occurs when an economic activity affects others. The externality can be positive as in the case of training and human capital improvement or it can be negative as in the case of environmental damage.

A. Bhattacharyya
Ernst and Young LLP, New York City, NY, USA

L. Kutlu
University of Texas Rio Grande Valley, Edinburg, TX, USA

R. C. Sickles (✉)
Department of Economics, Rice University, Houston, TX, USA
e-mail: rsickles@rice.edu

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_13

Government interventions that may lead to market distortions include controlled prices, taxes, trade controls, etc. For example, tariffs on imports increase the prices of relevant imports and their substitutes above their costs, insurances, and freight (cif) prices. The distortion, however, is not limited to price divergence in the imported goods. Since the domestic prices increase relative to the world prices, this affects exchange rates too. Hence, in some cases in order to determine the efficiency prices, we need to rely on an approach that considers macroeconomic factors as well.

In the presence of market failures, it would be sensible to identify the shadow value (efficient value) of relevant outputs or inputs. To this end, it is essential to understand the relationship between market prices and shadow prices. This may help policy makers to determine the direction in which the mix of outputs or inputs should change in order to enhance social welfare. For example, Grosskopf et al. (1999) compare market school district administrative and teaching salaries in Texas with their corresponding shadow prices. This enables them to determine whether the schools are under-utilizing or over-utilizing their administrators and teachers. Similarly, using plant-level data taken from Wisconsin coal-burning electric utility plants, Coggins and Swinton (1996) compare the prices paid for sulfur dioxide (SO_2) permits and the corresponding shadow prices. Swinton (1998) uses a similar comparison using plant-level data taken from Illinois, Minnesota, and Wisconsin. They find that the shadow prices are close to the permit prices. There are many other similar examples that we will briefly talk about later in the review.

It appears that a sensible starting point is considering an undistorted market where the market prices and shadow prices coincide. As Drèze and Stern (1990) argue, even in a perfectly competitive market the prices may be distorted, e.g., when the income distribution is not "optimal." Due to this reason, it is even possible to have shadow prices that are distinct from the perfectly competitive market prices. However, in many occasions, this aspect is ignored and the perfectly competitive market is assumed to be socially optimal. In Sect. 4, we present an approach that aims to control for macroeconomic factors and distributional disturbances.

In the case of non-market goods or bads, the price is not observed. Since the utility of a consumer depends on not only market goods but also non-market goods and bads, a social planner who cares about the social welfare should allocate some value to the non-market goods and bads. Shadow pricing methods, which we will provide a review, can be used to account for environmental factors or in general the non-market goods and bads.

In the next section, we talk about market prices and efficiency prices in the case of imperfect competition and present the welfare effects of pricing with market power. In Sect. 3, we summarize some of the widely used valuation methods that are used in valuation of non-market goods, services, or bads, which may cause an externality. In Sect. 4, we introduce a valuation approach for projects that can accommodate not only allocative efficiency viewpoint but also their impact on the growth and redistribution of income. Section 5 discusses identification of shadow prices using different approaches while the following section concludes the chapter.

2 Imperfect Competition, Market Power, and Welfare

2.1 Measures Related to Market Power

Imperfect competition is one of the most commonly encountered reasons for why market prices diverge from efficiency prices. The antitrust literature relates this divergence to market power, which is the ability of a firm or a group of firms to set prices above the efficiency prices (or competitive prices). The extent of price distortion critically depends on the market structure, i.e., characteristics of the market and firms.

Structure-conduct-performance paradigm generally uses market concentration measures such as Herfindahl-Hirschman index (HHI) to describe the market structures. The HHI, which is defined as the sum of squared market shares, gives some idea about the extent of welfare loss due to the price distortions and market power. One particular advantage of this approach is that the HHI can be calculated using the market share data only. However, this measure ignores many of the important characteristics and aspects of the market such as capacity constraints, dynamic factors, durability of product, price discrimination,[1] and substitutes. For example, a typical market with perishable goods can be modeled in a static setting whereas a market structure with durable goods requires a dynamic model. Being a static measure, the HHI may not be suitable in this context. Moreover, the HHI is market-specific and thus does not provide information about firm-specific distortions. Although market share data is relatively easier to find compared

[1]For details of advancements in the price discrimination literature, see Armstrong (2006) and Stole (2007).

to other market data, calculation of market share involves some conceptual difficulties related to definition of the market. This, however, is a common problem for market power studies. Finally, the HHI is not always positively related with the welfare. For example, let's start from a situation with two symmetric firms. Now, assume that one of these firms reduces its production costs. This would tend to increase the welfare and reduce prices charged to consumers. However, the HHI will increase. Therefore, the changes in the value of HHI may not always be in line with the changes in welfare.

Another widely used measure of market power is the Lerner (1934) index, which is defined as the ratio of price-marginal cost markup and price:

$$LI = \frac{P - MC}{P} \tag{1}$$

where P is the market price and MC is the marginal cost. The benchmark scenario for the Lerner index is perfect competition where price equals marginal cost, and thus, the Lerner index equals zero for the benchmark scenario. As the output price diverges from the efficiency price, i.e., marginal cost, the Lerner index increases and reaches its maximum value at the inverse of price elasticity of demand (in absolute value). Unlike the HHI, the Lerner index directly measures the price distortion that stems from imperfect competition. Moreover, it can be calculated as either a firm-specific or a market-specific measure of market power. The market-specific Lerner index is usually calculated as the market share weighted average of firm-specific Lerner index values.

Estimation of market power is an important issue to public policy makers and empirical industrial organization economists. Lerner index provides a simple way to address this issue as long as the marginal costs can be calculated. However, the usual assumption of price being at least as great as the marginal cost may not hold under certain market situations. Prices may be lower than marginal costs if firms either engage in price wars, or intentionally lower price of one product to promote sales of other similar products, or practice price discrimination, or if pricing of a product includes coupon discounts. Weiher et al. (2002) adopt a novel approach to overcome problems associated with estimation of Lerner index for US airlines, where prices can be lower than marginal costs. Since prices can possibly be zero for customers buying air tickets with frequent flyer miles, Weiher et al. (2002) use $\left(\frac{p-MC}{MC}\right)$ as a measure of market power instead of the usual Lerner index. This formulation allows them to put less weight on the below marginal cost prices, and averages of these normalized indices lead to more reasonable results in their study of US airlines.

Similar to the HHI, the conventional Lerner index assumes profit maximization in a static setting so that marginal revenue equals marginal cost. However, in a market characterized by dynamic factors, the price and production are determined intertemporally. If the current decisions of a firm involve a stock variable such as goodwill or knowledge or a level of quasi-fixed output, then the Lerner index needs to be adjusted to take these factors into account. Similarly, in the presence of an exhaustible or renewable resource, the conventional Lerner index needs to be adjusted. Pindyck (1985) proposes using what he calls full marginal cost (FMC), which is marginal cost plus competitive user cost, rather than marginal cost:

$$LI = \frac{P - FMC}{P} = \frac{P - (MC + \lambda)}{P} \tag{2}$$

where λ is the competitive user cost of one extra unit of cumulative production evaluated at the monopoly output path. Note that the user cost may depend on the extent of competition and other aspects of the market. Since a market power measure aims to reflect price distortions in comparison with competitive prices, the competitive user cost should be used as the correction term when calculating the FMC. Moreover, the competitive user cost must be calculated using the monopolist's output path just as marginal cost being evaluated at the monopoly output level when calculating the conventional Lerner index. Pindyck's (1985) market power measure ignores how the firms interact with each other, and thus, this measure is concerned with the measurement of potential market power.

Another case where the Lerner index needs to be interpreted carefully is when the firms have capacity constraints. With capacity constraints, price exceeds marginal costs (i.e., Lerner index is positive), and this indicates a welfare loss relative to perfect competition (without capacity constraint). But, if the capacity constraints are exogenous, then they are not under the control of the firms. Therefore, the deadweight loss should be calculated compared to perfect competition under capacity constraints, which indicates that the Lerner index needs to be adjusted to reflect this interpretation of deadweight loss. Puller (2007) suggests an adjusted Lerner index for markets where the firms have capacity constraints. In particular, he examines the market power of firms in the California electricity market. In this case, the adjusted Lerner index is the same as Eq. (2) except that λ equals the shadow cost of the capacity constraint. Since this shadow cost is not directly observed, it needs to be estimated along with the marginal cost. We will discuss this issue later in this section.

Even after adjusting for dynamic factors or capacity constraints, the Lerner index may not reflect price distortions precisely if a proper notion of marginal cost is not used. More precisely, the standard approaches for calculating the Lerner index implicitly assume that the firms are fully efficient. However, in reality, imperfect competition may lead to managerial inefficiency in both revenue and cost. The managerial inefficiency is present for a given production technology and can be improved if the firms do not waste resources and make optimal decisions in the production process. In practice, a common approach is estimating a cost function and calculating the marginal cost from the cost function parameter estimates. Using these marginal cost estimates and observed prices, the Lerner index is calculated. However, this does not reflect inefficiencies of firms in the Lerner index. Note that here we interpret the Lerner index as a measure of welfare loss for given production technologies in the market. Since the inefficiency reflects suboptimal outcome in the production process for given production technologies, calculation of the Lerner index needs to reflect such inefficiencies. In a static setting, Koetter et al. (2012) propose an efficiency adjusted measure of the Lerner index to overcome this issue. In a dynamic strategic framework where firms have repeated interactions, Kutlu and Sickles (2012) propose other efficiency adjusted Lerner index measures, but they only concentrate on inefficiency in cost. The Lerner index measure of Kutlu and Sickles (2012) is given by:

$$LI = \frac{P - EFMC}{P} = \frac{P - (EMC + \lambda)}{P} \tag{3}$$

where $EFMC = EMC + \lambda$ is the efficient FMC, *EMC* is the marginal cost for the full efficiency scenario, and λ is a term that adjusts for dynamic factors. Although they use different approaches, both Koetter et al. (2012) and Kutlu and Sickles (2012) calculate *EMC* from the stochastic cost frontier estimates. In contrast to these studies, Kutlu and Wang (2018) present a game theoretical model that estimates *EMC* directly.

All these Lerner index variations mentioned above require the marginal cost information, which is not readily available in most cases and needs to be estimated using a cost function model or other means. However, since the total cost data contains sensitive information for the firms, they may be reluctant to share this information. Even when the total cost data is available, it may not be available for the market of interest. For example, Kutlu and Sickles (2012) and Kutlu and Wang (2018) argue that the airline-specific total cost of the US airlines is available for the whole industry, but the route-specific total cost data is not available. Therefore, this poses some

issues when estimating route-specific marginal costs and Lerner indices for the airlines. Moreover, in the case where the firms have capacity constraints, the shadow cost of capacity is not available as well.

The conduct parameter (conjectural variations) method enables the estimation of marginal cost and an alternative market power measure, which is called conduct parameter, without using the total cost data. The conduct parameter is simply a demand elasticity adjusted counterpart of the Lerner index, and similar to the Lerner index, it can either be firm-specific or market-specific. Bresnahan (1989) and Perloff et al. (2007) are two good surveys on conduct parameter models. Some of the earlier examples of this approach include Gollop and Roberts (1979), Iwata (1974), Appelbaum (1982), Porter (1983), and Spiller and Favaro (1984).

The conduct parameter approach measures the market power of firms "as if" the firms have conjectures about other firms' strategies so that the equilibrium outcomes may not be supported by the standard market conditions: perfect completion, Nash equilibrium (in quantity or price), and joint profit maximization. For instance, in a Cournot model, the conjecture is that the firms will have zero reaction, i.e., conjecture is the Nash assumption in determining the equilibrium. Given the action (in this case, output) of other firms, each firm chooses its output optimally. Basically, conduct parameter approach assumes that firms may act as if they have more general types of reactions. Note that, in the conduct parameters method, the conjectures of firms refer to what firms do as a result of their expectations about other firms' behaviors, and it does not necessarily reflect what they believe will happen if they change their actions, e.g., quantities or prices. Based on this interpretation, one can consider the conduct parameter as an index that takes a continuum of values. Since the existing theories (e.g., perfect competition, Cournot competition, and joint profit maximization) are consistent with only a few of these potential values, some researchers may not be comfortable with the idea of conduct parameter taking a continuum of values. Hence, they would categorize the estimated conduct parameter using the competitive behavior of firms by using statistical tests (e.g., Bresnahan 1987).

Since the conduct parameter approach is based on game theoretical models, the researchers may add some structure to the model that describes the market structure in a market. In particular, capacity constraints (e.g., Puller 2007; Kutlu and Wang 2018), dynamic factors (e.g., Corts 1999; Puller 2009; Kutlu and Sickles 2012), managerial inefficiency (e.g., Koetter et al. 2012; Kutlu and Sickles 2012; Kutlu and Wang 2018), multi-output production (e.g., Berg and Kim 1998; O'Donnell et al. 2007; Kutlu

and Wang 2018), price discrimination (e.g., Graddy 1995; Kutlu 2017; Kutlu and Sickles 2017), and other characteristics of the market and firms can be incorporated to the game theoretical model, which describes the characteristics of the imperfect competition and market. In the literature, most conduct parameter models assume imperfectly competitive behavior by firms only in one side of the market, e.g., output market, and the other side of the market, e.g., input market, is assumed to be perfectly competitive. Hence, in general, these models only consider price distortions in output market but not in the input market. O'Donnell et al. (2007) present a general conduct parameter model that allows imperfect competition in both input and output markets.

Although the conduct parameter method relaxes the cost data requirement and allows more structural modeling, this does not come without a cost. In order to estimate a conduct parameter model, one needs to estimate a system of equations consisting of a demand function and supply relation that is derived from the first-order conditions of the structural game that the firms are playing. Hence, the required variables are the same as one would need for estimating a demand-supply system but with the exception that one needs to be more careful about identification. More precisely, if the researcher is not careful about the functional form choices, the marginal cost and conduct parameter may not be separately identified. For example, it may be possible to confuse competitive markets with high marginal cost and collusive markets with low marginal cost. Lau (1982) and Bresnahan (1982) present some conditions for identification in this setting. As argued by Bresnahan (1982), this identification problem can be solved by using general demand functions that the exogenous variables not only lead to parallel shifts but also change the demand slope by rotations. The simplest way to achieve this is including an interaction term with the quantity variable. However, Perloff and Shen (2012) illustrate that such rotations may cause some multicollinearity issues. Another approach that enables identification is assuming a constant marginal cost, which does not depend on quantity but may depend on other variables. For certain commonly used conduct parameter settings (Lau 1982), the conduct parameter and marginal cost can be separately identified if the inverse demand function $P(Q,Z)$, where Q is the quantity and Z is a vector of exogenous variables, is not a separable function of Z in the sense that we can write $P(Q,Z)=f(Q,h(Z))$ for some functions f and h. An alternative possibility is using the non-parametric structural identification approach in Brown (1983), Roehrig (1988), and Brown and Matzkin (1998). Another approach is modeling the conduct as a random variable and achieving the identification through distributional assumptions.

Orea and Steinbucks (2012) and Karakaplan and Kutlu (2019) achieve identification using such distributional assumptions and using econometric tools from the stochastic frontier literature. They use skewness of the distribution of conduct parameter in order to identify marginal cost and conduct parameter separately. This allows them to relax some of the strong functional form restrictions on the demand and marginal cost functions. Kumbhakar et al. (2012) propose another approach that estimates market powers of firms using the stochastic frontier approaches.

The conventional models for assessing market power assume either price or quantity competition to be the only endogenous variable. In reality, the degree of market power in the product market is likely to be related to input markets such as R&D, advertisement, finance, labor, capacity, and so on. A few recent studies investigate the influence of input markets on market power at the product market level. For example, Röller and Sickles (2000) examine whether the degree of market power at the product market is sensitive to capacity. They specify and estimate a two-step structural model in which firms make capacity decisions first and then decide the product-differentiated prices. In this framework, costs are endogenized through the first stage, which has important implications for the measurement of market power in the product market. In particular, Röller and Sickles (2000) specify a product-differentiated, price-setting game under the duopoly assumption, where each producer faces a demand of the form:

$$q_i\left(p_i, p_j, Z_i\right), i = 1, \ldots, N, \tag{4}$$

where N is the number of producers, q_i is the quantity demanded, p_i is a price index for producer i, p_j is a price index for competitor's prices, and Z_i is a vector of producer specific, exogenous factors affecting demand. While producers can affect costs only through changes in prices in the short-run, they can change the capital stock in the long run, thereby changing the long-run cost structure. Adopting a conjectural-variation framework, the first-order conditions of the two-stage profit maximization game in which producers purchase capital in stage 1 and decide prices in stage 2 can be written as:

$$\frac{p_i - MC(.)}{p_i} = \frac{1}{\eta_{ii} - \theta \frac{p_i}{p_j} \eta_{ij}} \tag{5}$$

where η_{ii} is the own price elasticity, η_{ij} is the cross-price elasticity, $MC(.)$ is the marginal cost based on the short-run cost structure, and the market

conduct parameter $\theta \equiv \partial p_j / \partial p_i$, represents the degree of coordination in a price-setting game. Based on this framework and profit-maximizing principle of firms, Röller and Sickles (2000) discuss estimation of the model, specification tests regarding the relevance of the sequential set-up for measuring market power and apply their method to analyze the European airline industry.

There are some theoretical examples (e.g., Rosenthal 1980; Stiglitz 1989; Bulow and Klemperer 1999) that suggest that more intense competition may lead to higher price-cost margins. Boone (2008a, b) proposes market power measures that are theoretically robust yet can be estimated using data sets that are similar to the ones that are used in estimating price-cost margins. In particular, Boone (2008a) proposes relative profit differences (*RPD*) measure and Boone (2008b) proposes the relative profits (*RP*) measure. The *RPD* measure is defined as follows. Let $\pi(n)$ denote the profit level of a firm with efficiency level n where a higher n value means higher efficiency. For three firms with efficiency levels $n'' > n' > n$, let:

$$RPD = \left(\pi\left(n''\right) - \pi(n)\right) / \left(\pi\left(n'\right) - \pi(n)\right) \tag{6}$$

be a variable representing *RPD*. Boone (2008a) argues that in models where a higher competition reallocates output from less efficient firms to more efficient firms, *RPD* increases in the extent of competition. Therefore, this measure covers a broad range of models. The relative profits measure is defined as follows. For two firms with efficiency levels $n' > n$, let:

$$PD = \pi\left(n'\right) / \pi(n) \tag{7}$$

be a variable representing profit differences. This measure is a robust market power measure as well.

2.2 Welfare Analysis

Having discussed the market power aspect of pricing inputs and outputs, it is imperative that we look into the welfare effects of such pricing. The conventional argument against market power evolves around the fact that by charging a price that is higher than the marginal cost, a firm is able to grab higher surplus, leaving the consumers with a lower surplus compared to the competitive market outcomes. However, the gain in producer surplus is often not big enough to compensate for the loss in the consumer

surplus, unless the producer employs perfect price discrimination. Thus, in the presence of market power, it is likely that the market outcome will be inefficient in terms of total surplus maximization and the society will experience welfare loss. The degree of welfare loss depends not only on the market power, i.e., the extent to which a firm is able to raise price above the marginal cost, but also on the elasticity of demand and size of the market.

Inefficiency of a non-competitive market is rooted in the inequality between price and the marginal cost of production (after factoring out restrictions that firms face in a suitable way). As mentioned in the previous section, one must consider an adjusted benchmark while identifying inefficiency of a non-competitive market in the presence of exogenous constraints. Otherwise, one may end up with an upward bias in the measured inefficiency. However, in the absence of any exogenous constraints, the difference between price and marginal cost is an indicator of the divergence between the marginal benefit to consumers and the marginal costs to producers. For a given technology (and cost) of production, such divergence leads to inefficient allocation of resources and static welfare loss for the society. The social cost of misallocation due to the presence of extreme market power like monopoly can be approximated by the well-known welfare triangle showing the difference between gain in producer surplus and loss in consumer surplus, when price is higher than the marginal cost. Prominent empirical research in this regard includes Harberger (1954) and Rhoades (1982). Using differences among profit rates in the US manufacturing industries, Harberger (1954) measures the possible increase in social welfare by eliminating monopolistic resource allocation. Rhoades (1982) studies the US banking sector and calculates the deadweight loss due to monopoly in the US banking system. However, Formby and Layson (1982) suggest to use caution while analyzing the relationship between market power as measured by the Lerner index or profit rates and allocative inefficiency. They find that under conditions of linear and constant price elasticity of demand functions, changes in monopoly power, as measured by the Lerner index or profit rates, are not adequate to predict changes in the allocative inefficiency.

The lack of competitiveness in a market is also likely to be associated with lower productive efficiency through wastage of resources and managerial efforts, which in turn may have crucial welfare implications. The study by Good et al. (1993) is worth noting in this regard. They discuss welfare implications for the US and European airlines by measuring changes in productive efficiency and market power due to liberalization. Focusing on relative efficiency scores defined by a stochastic production function frontier for

selected US carriers over the period 1976–1986, they find a clear evidence of convergence toward a common efficiency standard under deregulation for US carriers. However, European carriers that did not enjoy deregulation to a similar extent suffered from low efficiency and associated costs during the period. To identify potential welfare gain from deregulation for European airlines, Good et al. (1993) estimate the existing market power and compare it with the simulated effects of increased competition due to deregulation in a product-differentiated, price-setting game framework.

The argument in favor of privatization also stems from the fact that it is likely to increase operating efficiency and performance of economic units, thereby improving economic welfare. Several countries implemented privatization in different sectors of the economy over time to improve economic performance. While studying the economic impacts of privatization of the electricity company in Sub-Saharan Africa, Plane (1999) finds substantial evidence in support of improved productive efficiency, total factor productivity gain, and a reduction in the relative price of electricity, as a result of which, consumers are the main beneficiaries of privatization.

The presence of market power may also impose cost inefficiency in production systems. Possible reasons for such inefficiency include lack of managerial effort in cost minimization, following objectives other than profit maximization and utilizing resources for unproductive purposes like maintaining and gaining market power. Hicks (1935) identifies the lack in managerial effort in maximizing operating efficiency in the presence of extreme market power like monopoly, as the "quiet life" effect of market power. Empirical evidence suggests that the cost of inefficiency due to slack management may exceed the social loss from mispricing. While studying US commercial banks, Berger and Hannan (1998) find strong evidence for poor cost efficiency of banks in more concentrated markets. They also point to the fact that the efficiency cost of market concentration for US banks may outweigh the loss in social welfare arising from mispricing. On the contrary, in the specific case of European banking sector, Maudos and Guevara (2007) find welfare gains associated with a reduction of market power to be greater than the loss of cost efficiency—rejecting the "quiet life" hypotheses.

Finally, it is worth noticing that while the presence of market power may be associated with inefficiency and welfare loss, a producer with market power may also provide better-quality products and spread information through advertising, which in turn may contribute to the gain in economic well-being of consumers. A major difficulty in assessing welfare consequences of market power further arises from the fact that in more globalized economies with segmented markets and differentiated products, it is not straightforward to precisely define a market.

3 Externalities and Non-market Valuation

Besides imperfect competition, externalities are other commonly encountered reasons for why market prices diverge from efficiency prices. In a market, an externality is present when the production or consumption of a good leads to an indirect effect on a utility function, a production function, or a consumption set. Here, indirect refers to any effect created by an economic agent that is affecting another agent, where the effect is not transmitted through prices. In many occasions, the indirect effect is due to a produced output or used input that may not have a market value. For example, if a production process involves carbon dioxide (CO_2) emission that results in climate change, this would cause a negative effect on the society. However, no individual producer would try reducing the CO_2 levels unless there is some cost imposed on the emission levels or another mechanism that restricts emission. Another example is the production of public goods and services that provide benefits to the society, i.e., positive externality. Hence, the externality leads to distortions in the market prices that lead to deviations from efficiency prices unless it is somehow internalized. One potential way to internalize the externality is creating markets for non-market value inputs and outputs, which requires determining their values. The literature on valuation of non-market goods is vast. Hence, we only provide a broad summary of this literature. Additional summaries of literature on both theory and methods are given by Freeman (1979, 2003). Broadly speaking, there are two types of general approaches that are used in determining the valuation of goods in the presence of externalities—approaches based on technical relationships and behavioral (link) approaches, which rely on responses or observed behaviors. For the technical relationship approaches, we consider the damage function approach and distance function related approaches. For the behavioral approaches, we consider travel cost approach, hedonic pricing approach, and contingent valuation approach. While this list is not exhaustive, it covers some of the most widely used approaches in the literature.

3.1 Damage Function Approach

A procedure that belongs to the first group is the expected damage function approach. This method assumes a functional relationship between the good (bad) and expected social damage from decreasing (increasing) the amount of the good (bad). This approach is commonly used in risk

analysis and health economics. Rose (1990) (airline safety performance), Michener and Tighe (1992) (highway fatalities), Olson (2004) (drug safety), and Winkelmann (2003) (incidence of diseases and accident rates) exemplify some studies that use this approach in the risk analysis context. In general, the expected damage function approach can be used to measure the value of a good or a service (bad) that provides benefit in terms of decreasing (increasing) the probability and severity of some economic negative effect by the reduction in the expected damage. In an early application of this approach in the context of non-market valuation, Farber (1987) estimates the value of gulf coast wetlands due to its role of protection from wind damage to property that results from hurricanes. Obviously, the wetlands are non-market inputs, and thus, we cannot observe its value directly. The methodology aims to estimate a hurricane damage function in which wetlands moved by storms are a variable that determines the damage. He calculates the expected marginal damage from winds due to loss of the wetlands using the historic hurricane probabilities. Another example that uses the damage function approach is Barbier (2007) who measures the effects of mangrove forests on tsunami damages in Thailand. While the applications directly model the damage function, the starting point of this approach is the compensation surplus approach used for valuing a quantity or quantity change in non-market goods or services. In this setting, the expected damage due to a change in the amount of non-market good or service is the integral of the marginal willingness to pay for services that protect from the damage (e.g., avoid storm damage). The approach is useful in many occasions, but it only concentrates on one aspect of incremental benefits at a time. Hence, evaluation of the full valuation of a non-market good/bad or service would be difficult as this requires considering all aspects.

3.2 Distance Function Approach

Another non-market valuation approach based on technical relationships is the distance function approach. When the data on inputs and outputs is available, this enables us to construct a production model through a distance function. The properties of distance functions enable us to calculate the shadow prices for the inputs or outputs of production, which can be used to assign values to non-market goods or services. Färe and Grosskopf (1990) model the technology using input distance functions to represent technology. They use the duality of input distance function and cost function to calculate the cost normalized shadow prices. Färe et al. (1993) model the

technology using output distance function, which can accommodate multiple outputs and allows weak disposability of undesirable outputs. They obtain the normalized shadow prices by applying a dual Shephard's lemma and convert this to absolute shadow prices by assuming that the shadow price of one marketable output equals its market price. Another related approach to calculate shadow values is using the directional distance functions developed by Chambers et al. (1996, 1998). Chung et al. (1997) is the first example that models goods and bads using the directional distance functions. Among others, Lee et al. (2002), Färe et al. (2005), Färe et al. (2006), and Cross et al. (2013) are examples that use the directional distance function approaches. In contrast to Shephard's (1953, 1970) distance functions, which are defined in terms of radial expansions to the frontier, the directional distance functions are defined in terms of directional expansions along a specified vector. The radial distance functions are special cases of the directional distance functions. The directional distance function approach allows non-proportional changes in outputs (and inputs). Moreover, this approach allows mixture of expansions and contractions for outputs. That is, while some outputs may be expanded, the others can be contracted. Although the choice of direction is left to the researcher, a common choice is the unit vector with negative signs for bads. The trade-off between the good and the bad outputs is not meaningful unless technical efficiency is removed by projecting on the frontier. The issue is that such projections are not unique as there are competing projection methods and we need to choose one of them. Moreover, a change in a bad and a good output as we move from one point on the frontier to another depends on the direction and the size of the change. Hence, for an inefficient point, a directional projection may be a more sensible choice as it lies between the bad-oriented and the good-oriented projections. However, this flexibility in the choice of direction vector raises some concerns. In particular, the estimates may be sensitive to the direction choice as illustrated by Vardanyan and Noh (2006). Moreover, unlike the directional distance functions, the conventional radial distance functions allow unit-free multiplicative changes in arguments. Therefore, these two approaches do not have a decisive winner and the choice depends on the particularity of the problem that a researcher wants to answer. Finally, a general concern about distance functions is that modeling goods and by-products in the same technology may not be sensible. Fernández et al. (2002), Førsund (2009), and Murty et al. (2012) raise this concern and suggest separating the technology of goods and by-product bads. For this purpose, Fernández et al. (2002) assume that two technologies are separable and Murty et al. (2012) use distinct technologies. Acknowledging these

issues, Bokusheva and Kumbhakar (2014) present an approach that models the technology by two functions. They use a single technology specification but allow good and bad outputs to be related via a hedonic function. They provide the shadow price of the bad (pollutant) under the assumption that the shadow price of the marketed output equals its market price. Another paper that utilizes hedonic functions in this context is Malikov et al. (2016), which models undesirable outputs via a hedonic output index. This ensures that pollutants are treaded as outputs with undesirable nature as opposed to inputs or frontier shifters. For this purpose, Malikov et al. (2016) use a radial input distance function generalized to allow an unobservable hedonic output index of desirable and undesirable outputs.

Finally, we finish our notes about distance function approach by some application examples from a variety of contexts. Färe et al. (1993) (effluents by paper and pulp mills), Coggins and Swinton(1996), Swinton (1998), Färe et al. (2005) (SO_2 emission), Hetemäki (1996) (sulfate pulp plants), and Aiken and Pasurka (2003) (SO_2 and PM-10 emissions) exemplify some studies that concentrate on undesirable outputs. Other examples for shadow price estimates include Färe et al. (2001) (characteristics of sites), Aiken (2006) (activity of recycling), and Cross et al. (2013) (vineyard acres by quality).

3.3 Travel Cost Approach

The travel cost approach is developed by Trice and Wood (1958) and Clawson (1959). A good review is Parsons (2017). This approach belongs to the group of behavioral approaches, which is based on revealed preferences. In the context of environment, this method relies on the complementarity of quality of a natural resource and its recreational use value (e.g., visiting a national forest or fishing at a lake). The idea is that as the quality of a natural resource (e.g., quality of water) changes, the demand for the natural resource shifts. The change in the consumer surplus can be used to determine the value associated with the incremental benefit. Hence, individuals' willingness to pay for the recreational activity is revealed by the number of trips that they make and where they choose to visit among the potential options. Two subcategories of the travel cost models are single-site models and random utility maximization models. The single-site models consider the travel cost as the price and work similar to a demand function where the total number of trips is treated as the quantity of demand. On the other hand, the random utility maximization models assume multiple choices for the

individuals where the random utility is maximized based on these choices. In the random utility model, the sites are characterized by their attributes and travel cost for reaching the site. By choosing sites, the individuals reveal their preferences. Prior to the random utility travel cost models, multiple sites models were introduced in a demand system (Burt and Brewer 1971; Cicchetti et al. 1976). The random utility models became popular around the 1980s and 1990s starting with the works of Bockstael et al. (1984, 1987) on beach use and Carson et al. (1987) on recreational fishing. Parsons and Kealy (1992) and Feather (1994) (choice set formation), Adamowicz (1994) (intertemporal decisions), Train (1998) (simulated probability and mixed logit), and Hauber and Parsons (2000) (nested logit) exemplify some earlier works and developments around this time period. Meanwhile, the single-cite models concentrated on relaxing some other aspects of the problem such as continuity assumption of number of trips variable. This is achieved by using limited dependent variable and count data models (e.g., Shaw 1988; Hellerstein 1991, 1992; Hellerstein and Mendelsohn 1993). More recently, instrumental variable approach to handle endogeneity in congestion (Timmins and Murdock 2007) and models for handling on-site sampling are introduced in the random utility framework.

In a standard single-site model, the demand function is represented as:

$$q_i = f(p_i, ps_i, z_i, y_i) \tag{8}$$

where q_i represents the number of trips, p_i is the trip cost or price, ps_i is a vector of trip costs or prices for substitute sites, z_i is a vector of individual characteristics, and y_i is the income of individual i. A common choice for the demand function is the log-linear form. Using this demand function, the consumer surplus difference between with and without quality change can be used as a measure for quality improvement.

The random utility models provide a better behavioral explanation compared to the single-site models with an expanse of being somewhat more complicated. The individuals are assumed to choose among a set of possible sites (e.g., beaches, camping areas, parks, rivers, etc.) for a trip. In its simplest form, the utility from visiting a site is assumed to be a function of trip cost, p_{ki}, and a vector of site attributes (quality), X_i:

$$U_{ik} = \alpha p_{ki} + \beta X_i + \varepsilon_{ki} \tag{9}$$

where α and β are parameters and ε_{ki} is an error term. The individual picks the site that gives the highest utility:

$$V_i = \max(U_{1i}, U_{2i}, \ldots, U_{Ki}) \tag{10}$$

where U_{ki} is the utility from site k and V_i is the trip utility of individual i from visiting their top preference. If the quality level (e.g., more clean water) of a site, say U_{1i}, changes so that the new utility becomes $V_i^* = \max(U_{1i}^*, U_{2i}, \ldots, U_{Ki})$, the compensation variation measure for the trip is given by:

$$w_i = \frac{\left(V_i^* - V_i\right)}{-\alpha}. \tag{11}$$

3.4 Hedonic Pricing Approach

Hedonic price method is another approach that belongs to the group of behavioral approaches, which is based on revealed preferences. In this approach, the goods are characterized by their attributes or characteristics. The market transactions do not directly reveal the values of each characteristic, and this method aims to derive the values attached to these different characteristics of the goods indirectly. Quigley (1982), Freeman (1995), Bockstael and McConnell (2007), Phaneuf and Requate (2016), and Taylor (2017) are some reviews on hedonic pricing. Some of the applications of hedonic methods in a variety of markets include Griliches (1961) (automobile industry), Ridker and Henning (1967), Boyle et al. (1999) (housing markets), Triplett (1984) (computers), Triplett (2004) (information technology products), Primont and Kokoski (1990) (medical field), Schwartz and Scafidi (2000) (university education), and Good et al. (2008) (airline industry). The hedonic price method goes as early as Waugh (1928), but the utility theoretic connections between consumer preferences and equilibrium price for non-market valuation are provided in Rosen (1974).

The hedonic analysis has two stages. The first stage involves estimation of the hedonic price function. The second stage uses the first stage price estimates and combines them with the individual characteristics to estimate demand or utility function parameters. However, due to data availability limitations, the second stage is not always implemented. We will concentrate on the first stage. A detailed discussion on the second stage is given by Taylor (2017).

In a standard hedonic price analysis, estimation of the first stage involves regressing the price on the characteristics variables. Although there is no general rule for functional form choice, using the linear model requires some compelling reasons as the price and quality variables are likely to have some non-linear relationship. Cropper et al. (1988) provide evidence in support of relatively simpler models such as semilog functional form. However,

Kuminoff et al. (2010) find evidence supporting the more flexible functional forms. Another concern in price function estimation is the identification of model parameters. In particular, if the price variable is simultaneously determined with a characteristic variable or a relevant variable is omitted, this leads to inconsistent parameter estimates. The simultaneity problem can easily be handled by an instrumental variables approach (Irwin and Bockstael 2001). A particular omitted variable problem in the housing market context is omitting a relevant spatial lag variable, which can be addressed by using spatial hedonic price models. Anselin and Lozano-Gracia (2009) and Brady and Irwin (2011) provide extensive reviews for spatial hedonic price models.

3.5 Contingent Valuation Approach

Contingent valuation approach is the final behavioral method that we consider, which is based on stated preferences. This approach estimates the price of a good or a service through a contingent valuation question that carefully describes a hypothetical market. Contingent valuation method is useful when the market prices are unreliable or unavailable. Mitchell and Carson (1989) is an earlier book that provides a detailed discussion on designing a contingent valuation study. Boyle (2017) is a good recent review on contingent valuation for practical applications of the method. Although the approach has been widely critiqued, it is used in practice such as in some legal cases. Recently, Kling et al. (2012) argued that having some numbers is likely to be better than no number. On the other hand, Hausman (2012) focuses on the issues related to hypothetical bias and discrepancy between willingness to pay and willingness to accept. Therefore, the debate is still not conclusive.

Boyle (2017) identifies the steps in conducting a contingent valuation study as follows: (1) Identifying the change in quantity or quality to be evaluated; (2) identifying whose values to be estimated; (3) selecting data collection mode; (4) deciding about the sample size; (5) designing the information component of the survey instrument; (6) designing the contingent valuation question; (7) designing auxiliary questions; (8) pretesting and implementing survey; (9) analyzing data; and (10) reporting the results.

First, the researcher has to decide not only what needs to be measured but also whether there are risks involved. For example, in the case where there is some uncertainty about contamination of a water source, the valuation method would concentrate on the willingness to pay for a reduction in probability of contamination. The choice of whether the study

would be based on individuals or households is important. Quiggin (1998) argues that if intra-household altruism does not exist or it is paternalistic, the aggregate measure of welfare is the same. Whereas Munro (2005) argues that this happens when the household incomes are pooled. Bateman and Munro (2009) and Lindhjem and Navrud (2011) illustrate that the values for individuals and households differ. Traditionally, the most widely used survey method has been by mail, but Internet surveys became popular recently due to its cost and convenience advantages. However, response rates for Internet surveys are relatively lower compared to the other means. The cost or response rates are not the only concerns when choosing a survey method. Boyle et al. (2016) find that the Internet-based surveys give 8% lower estimates for willingness to pay compared to the other methods of survey. An important aspect of these surveys is description of what is being valued. Bergstrom et al. (1990), Poe and Bishop (1999), and MacMillan et al. (2006) exemplify studies that illustrate sensitivity of the results to the information provided. Another important aspect of these surveys is the payment mechanism. The response formats in these contingent valuation questions include open-ended (direct statement of willingness to pay) (Hammack and Brown 1974), iterative bidding (bid increases if respondent says yes to a bid and decrease for a no) (Randall et al. 1974), payment-card (choose among possible willingness to pay options) (Mitchell and Carson 1989), or dichotomous choice (yes or no to a specified willingness to pay amount) (Bishop and Heberlein 1979) questions. Among these, dichotomous choice questions are most commonly used. Carson and Groves (2007) and Carson et al. (2014) present conceptual arguments for desirable properties of this type of questions.

4 Macroeconomic Valuation of Projects: LM Methodology

As mentioned in the introduction, even in the case of perfect competition, the prices may be distorted if the income distribution is not optimal. The early days of cost-benefit analysis literature aimed to assess projects based on not only allocative efficiency viewpoint but also their impact on the growth and redistribution of income. Both optimal growth and optimal income distribution are important factors that need to be considered when evaluating the value of projects as suboptimal growth or income distribution leads to welfare loss. Hence, Little and Mirrlees (1969, 1974) (LM) and UNIDO (UN Industrial Development Organization) (1972) develop approaches

that aim to address this objective. The approach of LM is subsequently extended by Squire and van der Tak (1975) and UK Overseas Development Administration (1988). Combining allocative efficiency, growth, and redistribution aspects requires a common measure, which may be aggregated into a single measure. The LM approach uses the world price as numeraire. This method converts the domestic prices to world prices by using the standard conversion factor. Note that this does not claim that the world prices are undistorted and reflect perfectly competitive prices. Rather, the world prices are used because they represent the conditions in which the economy can participate in world trade and they reflect comparative advantages. On the other hand, UNIDO (1972) uses the domestic price numeraire and converts domestic prices using the shadow exchange rate. The approaches of LM and UNIDO are similar in spirit but the LM approach is a more widely adopted methodology for shadow price estimation. Therefore, in this section, we concentrate on the LM approach. Further details can be found in the cited studies as well as in Chowdhury and Kirkpatrick (1994) and Asian Development Bank (2013).

The valuation of public projects requires prices for traded and non-traded goods. The LM approach determines the valuations of traded goods based on world prices, which reflect the opportunity costs to the country evaluating the project. This reflects the net benefit of a traded good. The non-traded goods are not traded internationally due to either an export ban or another reason. Since the traded goods are valued at world prices, the non-traded goods should be valued comparably. This is achieved by first estimating the marginal cost of production and converting input costs to world prices. The conversion involves decomposing the inputs into traded and non-traded inputs labor and land. Then, the non-traded land and labor prices are converted into world prices. This conversion process involves determining the traded goods that they substitute in domestic production. The world prices of these goods can be used in order to drive shadow prices for the non-tradable goods.

As mentioned earlier, shadow prices for traded goods are based on world prices. In particular, for imports cif and for exports fob (free on board) prices are used. The prices can be given in either foreign exchange terms or domestic currency values. The world prices need to be adjusted for the costs of internal transportation and distribution. Since the world price is intrinsically an abstract concept, it must be estimated. One challenging issue is that the goods are rarely homogenous. Moreover, the goods may be subject to different price discrimination practices, e.g., different unit prices for different amounts. Hence, it is impossible to avoid researcher's judgment when calculating the world price estimates.

A common way to calculate a shadow price for a non-tradable good is using a conversion factor, which is the ratio between the market price and the shadow price of the good. The shadow value for the relevant non-tradable good is calculated by multiplying the market price with the relevant conversion factor. Whenever the researcher does not have enough information about the non-tradable good or if the amount of the non-tradable good is small, the so-called standard conversion factor is used.

The development of semi-input-output method helped the consistent estimation of macroconversion factors. After identifying a set of primary factor inputs, primary inputs are given (exogenously or endogenously determined) values. Then, the economic price of a sector s (EP_s) is determined by a weighted average of conversion factors of primary inputs x into s:

$$EP_s = \sum\nolimits_x vx_s CF_x \tag{12}$$

$$CF_s = \frac{EP_s}{FP_s} \tag{13}$$

where vx_s is the value of primary input x into sector s, FP_s is the financial price value of s. This approach has the disadvantage of input-output systems as they employ fixed coefficients. However, it has the advantage of picking up both direct and indirect effects. For example, not only the direct employment effects but also the linkage employment effects from expansion of production are reflected.

Especially in economies with labor surplus, unskilled labor takes an importance place. In LM approach, the shadow price of unskilled labor is calculated using a separate conversion factor. If the production involves multiple goods, then the weighted mean of conversion factors for each output produced is applied to the market value of the opportunity cost of unskilled labor. The skilled labor shadow price is calculated by applying the standard conversion factor to the market wage.

As stated earlier, one of the aspects of LM approach is that it takes distributional issues into account as well. This is particularly important because the policy maker may not only be interested in allocative efficiency but also be interested in how the resources are distributed. LM approach considers two types of distributional issues. The first one is about distribution of output among members of the society with different incomes. The second one is about intertemporal distribution of resources. This involves deciding about which portion of a project's output will be saved and which portion of the project's output will be consumed. These procedures involve

assigning distributional weights, which in turn contribute to the calculation of the shadow prices. For example, the poor are given higher weights compared to the rich. Squire and van der Tak (1975) present this approach more formally and show how distributional weights can be fed into a variety of parameters. Ray (1984) formalizes many of the expressions by Squire and van der Tak (1975) further and explains the underlying welfare theory. In practice, however, the distributional weight approach is not applied without some concerns. The main issue is that many times the weights are based on value judgments. Harberger (1978) argues that the weighting scheme gives implausibly high/low weights to some groups. Some even argue that even equal weights are subjective (e.g., Brent 2006). The arbitrarily chosen weights may make the allocative efficiency less important than the distributional objectives. Hence, sometimes the analysis for allocative efficiency and distributional impact is made separately. Overall, the LM method provides us a macroperspective when evaluating valuations of projects and is a useful tool along with other valuation methods that we summarized in this short review.

5 Shadow Prices of Inputs and Outputs

Shadow prices are virtual prices that can be calculated as changes in the optimal value of an objective function for marginal relaxation in the constraint, in a constrained optimization framework. Inevitably, shadow prices are highly relevant in constrained output, revenue, profit maximization, and cost minimization problems faced by production units. These prices are primarily theoretical values, estimation of which can be useful when market prices do not exist or do not reflect the true value of products. There are several approaches for identifying and estimating measures related to shadow prices in the productivity literature. These approaches differ in their objective functions, nature of inputs and outputs, and methods of identification.

5.1 Shadow Prices Based on the Cost Function

One plausible approach for identifying shadow price measures for inputs is to focus on the dual profit function. Under the standard regularity production conditions, this approach allows one to identify the profit-maximizing output supply and input demand system by virtue of the Hotelling's lemma. The output supply and input demand functions then can be modified to

incorporate different types of inefficiency, which in turn reflect the relationship between the perceived and the actual market prices of inputs and outputs. For example, Lovell and Sickles (1983) use the dual profit function to model technology of a competitive profit-maximizing multi-product firm and estimate the ratio of perceived to actual price of inputs. This ratio reflects the systematic component of allocative inefficiency and plays a pivotal role in estimating the cost of the forgone profit due to inefficiency. Based on this approach, Sickles et al. (1986) study the US airline industry for allocative distortions during a period of regulatory transition.

In the presence of quasi-fixed inputs, the production technology can be modeled using a dual restricted (variable) cost function that allows for the existence of temporary disequilibrium (Sickles and Streitwieser 1998). Temporary disequilibrium may occur for unexpected demand shocks or changes in factor prices. Under the assumptions of exogenous input and output prices, the short-run variable cost function can be obtained as a solution of the minimization problem of a firm operating at full capacity:

$$\min \sum W_i X_i \text{ subject to } H(Y, X; T) = 0 \tag{14}$$

where H is the transformation function of the production technology, Y is the output, W represents input prices, and X represents the quantity of quasi-fixed inputs. The short-run variable cost function is then given by:

$$CV = G(Y, W, X; T) \tag{15}$$

where G is linearly homogeneous, non-decreasing, and concave in input prices; non-decreasing and convex in the levels of quasi-fixed inputs; and non-negative and non-decreasing in output. For example, G can be a non-homothetic translog function. Then, given exogenous input prices and by Shephard's lemma, the first-order conditions of the cost minimization problem yield the variable cost share (M_i) for variable input (X_i). For estimation purposes, the shadow share equation, $-\frac{\partial lnG}{\partial lnX_k} = \frac{Z_k X_k}{CV}$, can be added to the model where the shadow price, Z_k, is the real rate of return or *ex-post* value of the fixed input Z_k. The shadow price can be derived as the residual between revenues and variable costs. Since the effects of economic optimization are incorporated in the shadow value equations, they can be used in the system of estimating equations. The long-run cost function can also be obtained from the restricted cost function as $C = H(W, Y, Z^*)$ where $Z_k^* = -\frac{\partial G}{\partial X_k}$. Sickles and Streitwieser (1998) apply their model and methodology to study the interstate natural gas transmission industry in the USA.

A production technology may be subject to inherent complexities, constraints, and distortions that are needed to be integrated into the optimization problems of firms. Good et al. (1991) formulate a multiple output technology in which the choice of production technique is an endogenous decision. They employ the concept of virtual prices in their modeling to estimate technology that corresponds to efficient resource allocation. They also discuss estimation of parameters that explain the divergence between virtual and observed prices and apply their method to analyze the US airline industry.

The institutional constraints and policy environment can substantially affect the relative input price in unobserved ways, resulting in a divergence between the relative market price and the relative shadow price. Extent of this divergence measures the relative price efficiency. Getachew and Sickles (2007) estimate the divergence of the relative market price form the relative shadow price using a generalized cost function approach. The first-order conditions for a standard neoclassical problem of cost minimization subject to an output constraint yield the equality between the marginal rate of technical substitution (MRTS) and the ratio of market price of inputs. However, in the presence of additional constraints due to the policy environment, the optimal allocation of inputs that minimize cost requires the equality between the MRTS and the ratio of shadow or effective prices. Thus, a firm's cost minimization problem in the presence of additional restrictions can be given as:

$$\min_X C = P'X \text{ s.t.} f(X) \leq Q \text{ and } R(P, X; \varphi) \leq 0 \tag{16}$$

where P and X are $h \times 1$ vectors of price and quantity of inputs, respectively, $f(X)$ is a well-behaved production function, Q is output, $R(.)$ is an R_C-dimensional function representing additional constraints, and φ is a vector of parameters. The first-order conditions for cost minimization become

$$\frac{f_i}{f_j} = \frac{P_i + \sum_{r=1}^{R_C} \lambda_r \partial R_r / \partial X_i}{P_j + \sum_{r=1}^{R_C} \lambda_r \partial R_r / \partial X_i} = \frac{P_i^e}{P_j^e}, i \neq j = 1 \ldots h \tag{17}$$

The parameters of the unobservable shadow prices then can be estimated using a first-order Taylor series approximation to a general shadow price function $g_i(P_i)$ such that $g_i(0) = 0$ and $\frac{\partial g_i(P_i)}{\partial P_i} \geq 0$. One way to approximate these shadow prices (Lau and Yotopoulos 1971; Atkinson and Halverson 1984) is to consider:

$$P_i^e = k_i P_i, i = 1 \ldots h \tag{18}$$

where k_i is an input-specific factor of proportionality, the value of which informs us about the price efficiency of inputs. The shadow cost function in this case is given by:

$$C^S = C^S(kP, Q) \tag{19}$$

Using the logarithmic differentiation and Shephard's lemma, one can derive the input demand functions from the shadow cost function and hence can derive the actual cost function and share equations. In particular, the demand for factor i is:

$$X_i = \frac{M_i^S C^S}{k_i P_i}, i = 1 \ldots h \tag{20}$$

where M_i^S is the shadow share of factor i. Thus, the actual cost function $\left(C^A\right)$ and the actual share equation for input i are derived as $C^A = C^S \sum_{i=1}^{h} \frac{M_i^S}{k_i}$ and $M_i^A = \frac{X_i P_i}{C^A}$, respectively. Getachew and Sickles (2007) use this econometric model to analyze the Egyptian private manufacturing sector.

The regulatory constraints are likely to have major implications for the productivity and resource costs of production systems. For example, regulations regarding capital requirements affect resource costs in banking systems. Duygun et al. (2015) discuss measurement of shadow returns on equity associated with regulatory capital constraints on emerging economy banking systems. They model the cost function by incorporating regulatory constraints and measure productivity cost of changes in the regulatory capital requirements by measuring shadow price of the equity capital over time. In particular, in the presence of regulated equity-asset ratio in banking systems, they model the parametric frontier dual-cost function as:

$$c(y, w, r_0, t) + w_0 z_0 = \min_x \left\{ w'x + w_0 z_0 : F(x, z_0, y, t) = 0, z_0 = r_0 y \right\} \tag{21}$$

where x, w, and y are vectors of variable inputs, input prices, and output, respectively, and z_0 is a particular input that is either fixed in the short run, or required in a fixed ratio to output, but variable in the long run. Price of z_0 is w_0. The transformation function $F\left(x, z_0, y, t\right)$ is the efficient boundary of the technology set. Assuming weak disposability and applying the envelop theorem showing the relationship between the long-run and short-run total cost, they derive the shadow price interpretation of the target equity capital ratio in terms of the shadow share of equity costs to total expenses as:

$$-\left[\frac{\partial c\left(y, w, r_0^*, t\right)}{\partial lnr_0}\right] = \left(w_0 y\right)\left(\frac{r_0}{C}\right) = \left(\frac{w_0 z_0}{C}\right). \tag{22}$$

Applying this model, Duygun et al. (2015) confirm the importance of regulated equity capital as a constraint on cost minimizing behavior of banks in emerging economies.

The literature in this area has expanded to incorporate dynamic production and cost models as well. Captain et al. (2007) introduce a dynamic structural model to simulate the optimal levels of operational variables and identify sources of forgone profit. Using Euler equations derived from the first-order conditions of a dynamic value function maximization problem along with demand and cost equations, they simulate operating behavior of production units. They apply their model using data from the European airline industry to identify inefficiency in airlines by comparing the simulation results with the actual data and identify several sources of forgone profit like suboptimal network size. The methodology and modeling approach used in Captain et al. (2007) can be used to analyze potential impacts of economic policies in other setting as well.

While the shadow cost minimization based on shadow prices is widely used in the literature for identifying shadow values of inputs, an alternative approach is to use a shadow distance system. The shadow distance system can be estimated both in the static framework and in the dynamic framework, which accounts for adjustment costs of inputs. Atkinson and Cornwell (2011) discuss minimization of shadow costs of production in a dynamic framework using input distance function. Their formulation is based on the idea that shadow input quantities are likely to differ from actual input quantities, resulting in an inequality between the marginal rate of substitution and input price ratios. Divergence between the shadow and actual input quantities can occur due to policy regulations, contractual obligations, or shortage. Further, the production process may involve adjustment costs in terms of reduced output during the initial testing phase of a new capital good or training period of a newly hired worker. In this framework, they estimate the shadow costs by estimating a set of equations including the first-order conditions from the short-run shadow cost minimization problem for the variable shadow input quantities, a set of Euler equations derived from subsequent shadow cost minimization with respect to the quasi-fixed inputs, and the input distance function expressed in terms of shadow quantities.

Tsionas et al. (2015) further expand the literature by proposing estimation methods for a flexible system of input distance function in the presence of endogeneity of inputs. In their study, they discuss computation of the cost of allocative inefficiency, which is defined as the predicted difference between the actual and the frontier cost and is computed as a fraction of the predicted frontier cost. They apply their model and method to analyze production of Norwegian dairy firms.

Based on the standard economic model of shadow cost minimizing behavior of firms, it seems a natural choice to use shadow input quantities while analyzing cost minimizing behavior of firms. However, this approach involves significant challenges in terms of estimation. Coelli et al. (2008) propose a model based on shadow input prices in a similar framework and identify allocative inefficiencies in terms of shadow input prices. They also apply their method to a panel on US electricity generation firms.

5.2 Shadow Prices Based on the Directed Distance Function

Many production technologies produce undesirable or "bad" outputs along with desirable or "good" outputs. Some examples of undesirable outputs include the environmental degradation associated with use of pesticides in farming and greenhouse gas emission from industrial production technologies. It is logical to adjust producer performance based on shadow values of the undesirable outputs produced as a by-product of desired outputs. However, the undesirable outputs are often non-marketable, and thus, the valuation of such outputs is not straightforward. Often policy regulations are imposed to restrict the ability of producers to costlessly dispose of undesirable outputs. These regulations involve abatement of pollutants. There is an associated opportunity cost to the abatement process, which is the forgone marketable output. One possible approach of measuring shadow price of undesirable outputs is to rely on the data on abatement cost. The problem with this approach is that the data on abatement cost is likely to be subject to a wide range of errors.

An alternative approach is to estimate an output distance function which is dual to the revenue function (Shephard 1970). Then, by dual Shephard's lemma, the output distance function yields the revenue deflated shadow prices of all outputs, including undesirable outputs. In particular, the output distance function, as introduced by Shephard (1970), is given by:

$$D_0(x,u) = \inf\left\{\theta : \left(\frac{u}{\theta}\right) \in P(x)\right\} \tag{23}$$

where $x \in \mathcal{R}_+^N$ is the input vector, $u \in \mathcal{R}_+^M$ is the output vector, and $P(x) = \left\{u \in \mathcal{R}_+^M : x \text{ can produce } u\right\}$ represents the convex output set. Under the assumption that the technology satisfies the standard properties and axioms (Shephard 1970; Färe 1988), $P(x)$ satisfies weak disposability of outputs, meaning reduction of an undesirable output can be achieved by simultaneously reducing some desirable output(s).

Färe et al. (1993) discuss the process to retrieve output shadow prices from the following duality relationships between the revenue function and the output distance function:

$$R(x,r) = \sup_u \{ru : D_0(x,u) \leq 1\} \tag{24}$$

$$D_0(x,u) = \sup_r \{ru : R(x,r) \leq 1\}, \tag{25}$$

where ru is the inner product of output price and quantity vectors, $r \neq 0$. Assuming the revenue and output distance functions are differentiable, and the output distance function is linearly homogeneous in outputs, the first-order conditions of the Lagrange problem can easily be written as:

$$r = R(x,r) \cdot \nabla_u D_0(x,u) \tag{26}$$

Further, from the second duality relationship, we have:

$$D_0(x,u) = r^*(x,u)u \tag{27}$$

where $r^*(x,u)$ is the revenue maximizing output price vector from the second duality condition. Then, by Shephard's dual lemma:

$$\nabla_u D_0(x,u) = r^*(x,u) \tag{28}$$

and therefore:

$$r = R(x,r) \cdot r^*(x,u). \tag{29}$$

The $r^*(x,u)$ term can be interpreted as a vector of normalized or revenue deflated output shadow prices. Since $R(x,r)$ depends on the vector of shadow prices r, for identification purposes one needs to assume that one observed output price equals its absolute shadow price. This assumption can easily be justified for a desired output, which is observable and market-determined price. This approach is straightforward to implement with a suitable parameterization of the output distance function. Färe et al. (1993)

point out that shadow prices retrieved by this approach "reflect the trade-off between desirable and undesirable outputs *at the actual mix of outputs* which may or may not be consistent with the maximum allowable under regulation." They showcase their method on a sample of paper and pulp mills in the USA.

In a recent working paper, Färe et al. (2015) use an input distance function and dual Shephard's lemma to derive shadow prices and use them to construct an imputed price index. Assuming that a good is endowed with $z = (z_1, \ldots, z_N)$ characteristics that generate a value $p \geq\geq 0$, they model the input correspondence as:

$$L(p) = \left\{ z \in \mathcal{R}_+^N : z \,\text{generates}\ \text{value}\, p \right\}, p \geq 0. \tag{30}$$

With the help of Shephard's (1953) input distance function and some mild assumptions on $L(p)$, they discuss a complete characterization of the input correspondence as:

$$D_i(p, z) \geq 1 \Leftrightarrow z \in L(p). \tag{31}$$

Then, the cost function, which is the dual to the input distance, can be given as:

$$C(p, w) = \min\{wz : z \in L(p)\} \tag{32}$$

where $w \in \mathcal{R}_+^N$ are the unknown prices of characteristics. Using the duality between $D(p, z)$ and $C(p, w)$, the shadow price vector can be obtained as:

$$w^s = \frac{p \cdot \nabla_z D_i(p, z)}{D_i(p, z)}. \tag{33}$$

Färe et al. (2015) illustrate this method by constructing property price indices for houses in Netherlands. They also point out that this method avoids the multicollinearity problem associated with traditional hedonic regression.

Over the last two decades, the issues related to productivity growth and environmental quality have drawn a great deal of attention from economists. It is more important for production processes that experience substantial production of undesirable output like carbon dioxide and other greenhouse gases in the course of producing desirable outputs. The traditional productivity indices assume that undesirable outputs, if any, are freely disposable. However, that is a very strong assumption to be imposed on the technology and is often violated in reality. When undesirable outputs are produced as by-products of desirable outputs, it is reasonable to assume

weak disposability of outputs. The weak disposability of output implies that a reduction in undesirable outputs can only be achieved by the reduction of desirable outputs, given fixed input levels.

Several studies (Chung et al. 1997; Boyd et al. 1999) focus on the construction of productivity indices in the presence of both "good" and "bad" outputs. The study by Jeon and Sickles (2004) is notably relevant in this regard. They use the directional distance function method to construct the Malmquist and Malmquist-Luenberger productivity indices under the assumption of weak disposability of undesirable outputs. While exploring a sample of OECD and Asian countries in their study, they discuss computation of incremental costs of pollution abatement. More specifically, they choose the direction vectors for the pollutant—carbon dioxide levels that are not freely disposable, derive the production frontier using the specific restrictions on carbon dioxide emissions, and calculate incremental costs by dividing the change in the frontier value of GDP under the assumption of free disposability by the corresponding frontier level of carbon dioxide emissions. The incremental costs of pollution abatement can give us a fair idea about the shadow values of pollution control and prices of pollution permits.

Undesirable output may be generated in other production systems as well. For example, banking services produce nonperforming loans that are not desired. While finding shadow prices of bank equity capital using parametric forms of directional distance functions, Hasannasab et al. (2018) consider deposits and borrowed funds as inputs to produce desirable outputs loans and leases along with undesirable output nonperforming loans. Since reducing undesirable output is costly, they assume undesirable and desirable outputs satisfy joint weak disposability while desirable inputs and outputs satisfy strong disposability. Accordingly, they obtain shadow prices using the estimated distance functions via the Lagrangian method. In the process, they use different pricing rules based on differently oriented distance functions that are associated with different economic optimization criteria like cost minimization, revenue maximization, and profit maximization.

6 Concluding Remarks

In this chapter, we discuss pricing methods that are adopted when the competitive or socially efficient prices are not established because of either market imperfections or externalities. We also discuss several shadow pricing methods and their implications when the market price is not observed or

when the commodity is not marketable. However, the degree of price distortion due to market imperfection is not easy to measure, and the standard indices like the Lerner index need adjustment for dynamic factors, capacity constraints, and inefficiency of the production system that can affect the cost of production. Further, data for estimating market power and hence the degree of price distortion may not be readily available, and the researcher may need to modify the relevant methodology accordingly. Similarly, in the presence of externalities, market prices are likely to be far away from efficient prices, unless the external effects are accounted for in the pricing methods. This is more relevant when a production system produces undesirable outputs along with the desired ones. There are different methods for identifying and internalizing such external effects. In this chapter, we discuss several directions based on the most recent literature for dealing with these issues.

The literature on the estimation of shadow prices and their efficiency implications has expanded vastly in the last three decades. We discuss several approaches in this regard based on different objective functions and the nature of inputs and outputs. The pricing methods are not only important from microeconomic perspectives but also from macroeconomic perspectives, especially for international trade, growth, and distribution. The welfare implications of pricing under different circumstances are also crucial for policy makers. While there is an apparent conflict between the producers' and consumers' interests, factors like advertising or quality improvement for maintaining market power positively affect both groups. Since total welfare is influenced by both producer and consumer surplus, it is not straightforward to measure the welfare impacts of different pricing policies. Though some researchers have ventured into measuring welfare impacts in this regard, as is discussed in the chapter, it is still an open area of research, both from the microeconomic and macroeconomic perspectives.

References

Adamowicz, W.L. 1994. Habit formation and variety seeking in a discrete choice model of recreation demand. *Journal of Agricultural and Resource Economics* 19: 19–31.

Aiken, D.V. 2006. Application of the distance function to nonmarket valuation of environmental goods and services: An illustrative example. *Ecological Economics* 60: 168–175.

Aiken, D.V., and C.A. Pasurka Jr. 2003. Adjusting the measurement of U.S. manufacturing productivity for air pollution emissions control. *Resource and Energy Economics* 25: 329–351.

Anselin, L., and N. Lozano-Gracia. 2009. Spatial hedonic models. In *Palgrave handbook of econometrics: Volume 2, applied econometrics*, ed. T. Mills and K. Patterson. Basingstoke: Palgrave Macmillan.

Appelbaum, E. 1982. The estimation of the degree of oligopoly power. *Journal of Econometrics* 19: 287–299.

Armstrong, M. 2006. Recent developments in the economics of price discrimination. In *Advances in economics and econometrics: Theory and applications, ninth world congress of the econometric society*, ed. R. Blundell, W. Newey, and T. Persson. Cambridge, UK: Cambridge University Press.

Asian Development Bank. 2013. *Cost-benefit analysis for development: A practical guide*.

Atkinson, S.E., and C. Cornwell. 2011. Estimation of allocative inefficiency and productivity growth with dynamic adjustment costs. *Econometric Reviews* 30: 337–357.

Atkinson, S.E., and R. Halvorsen. 1984. Parametric efficiency tests, economies of scale, and US electric power generation. *International Economic Review* 25: 647–662.

Barbier, E.B. 2007. Valuing ecosystem services as productive inputs. *Economic Policy* 22: 177–229.

Bateman, I.J., and A. Munro. 2009. Household versus individual valuation: What's the difference? *Environmental & Resource Economics* 43: 119–135.

Berg, S.A., and M. Kim. 1998. Banks as multioutput oligopolies: An empirical evaluation of the retail and corporate banking markets. *Journal of Money, Credit and Banking* 30: 135–153.

Berger, A.N., and T.H. Hannan. 1998. The efficiency cost of market power in the banking industry: A test of the "quiet life" and related hypotheses. *Review of Economics and Statistics* 80: 454–465.

Bergstrom, J.C., J.R. Stoll, and A. Randall. 1990. The impact of information on environmental commodity valuation decisions. *American Journal of Agricultural Economics* 72: 614–621.

Bishop, R.C., and T.A. Heberlein. 1979. Measuring values of extra-market goods: Are indirect measures biased? *American Journal of Agricultural Economics* 61: 926–930.

Bockstael, N.E., and K.E. McConnell. 2007. Environmental and resource valuation with revealed preferences: A theoretical guide to empirical models. In *The economics of non-market goods and resources series*, vol. 7, ed. I.J. Bateman. Dordrecht, The Netherlands: Springer.

Bockstael, N.E., W.M. Hanemann, and C.L. Kling. 1987. Estimating the value of water quality improvements in a recreational demand framework. *Water Resources Research* 23: 951–960.

Bockstael, N.E., W.M. Hanemann, and I.E. Strand. 1984. Measuring the benefits of water quality improvements using recreation demand models. Report presented to the U.S. Environmental Protection Agency. College Park: University of Maryland.

Bokusheva, R., and S.C. Kumbhakar. 2014. A distance function model with good and bad outputs, 2014 International Congress, August 26–29, Ljubljana, Slovenia 182765, European Association of Agricultural Economists.

Boone, J. 2008a. A new way to measure competition. *Economic Journal* 118: 1245–1261.

Boone, J. 2008b. Competition: Theoretical parameterizations and empirical measures. *Journal of Institutional and Theoretical Economics* 164: 587–611.

Boyd, G., R. Färe, and S. Grosskopf. 1999. Productivity growth with CO_2 as an undesirable output, mimeo.

Boyle, K.J. 2017. Contingent valuation in practice, A primer on nonmarket valuation. In *The economics of non-market goods and resources*, vol. 13, 83–131. Dordrecht: Springer.

Boyle, K.J., P.J. Poor, and L.O. Taylor. 1999. Estimating the demand for protecting freshwater lakes from eutrophication. *American Journal of Agricultural Economics* 81: 1118–1122.

Boyle, K.J., M. Morrison, D.H. MacDonald, R. Duncan, and J. Rose. 2016. Investigating Internet and mail implementation of stated-preference surveys while controlling for differences in sample frames. *Environmental & Resource Economics* 64: 401–419.

Brady, M., and E. Irwin. 2011. Accounting for spatial effects in economic models of land use: Recent developments and challenges ahead. *Environmental & Resource Economics* 48: 487–509.

Brent, R. 2006. *Applied cost-benefit analysis*. Cheltenham: Edward Elgar.

Bresnahan, T.F. 1982. The oligopoly solution concept is identified. *Economics Letters* 10: 87–92.

Bresnahan, T.F. 1987. Competition and collusion in the American automobile oligopoly: The 1955 price war. *Journal of Industrial Economics* 35: 457–482.

Bresnahan, T.F. 1989. Empirical studies of industries with market power. In *The handbook of industrial organization*, vol. 2, ed. Richard Schmalensee and Robert D. Willig, 1011–1057. Amsterdam: North-Holland.

Brown, B. 1983. The identification problem in systems nonlinear in the variables. *Econometrica* 51: 175–196.

Brown, D.J., and R.L. Matzkin. 1998. Estimation of nonparametric functions in simultaneous equations models with application to consumer demand. Working paper, Yale University.

Bulow, J., and P. Klemperer. 1999. Prices and the winner's curse. *RAND Journal of Economics* 33: 1–21.

Burt, O.R., and D. Brewer. 1971. Estimation of net social benefits from outdoor recreation. *Econometrica* 39: 813–827.

Captain, P.F., D.H. Good, R.C. Sickles, and A. Ayyar. 2007. What if the European airline industry had deregulated in 1979? *A Counterfactual Dynamic Simulation, the Economics of Airline Institutions* 2: 125–146.

Carson, R.T., and T. Groves. 2007. Incentive and information properties of preference questions. *Environmental & Resource Economics* 37: 181–210.

Carson, R.T., T. Groves, and J.A. List. 2014. Consequentiality: A theoretical and experimental exploration of a single binary choice. *Journal of the Association of Environmental and Resource Economists* 1: 171–207.

Carson, R.T., W.M. Hanemann, and T.C. Wegge. 1987. Southcentral Alaska Sport Fishing Study. Report prepared by Jones and Stokes Associates for the Alaska Department of Fish and Game. Anchorage, AK.

Chambers, R.G., Y. Chung, and R. Färe. 1996. Benefit and distance functions. *Journal of Economic Theory* 70: 407–419.

Chambers, R.G., Y. Chung, and R. Färe. 1998. Profit, directional distance functions and Nerlovian efficiency. *Journal of Optimization Theory and Applications* 95: 351–364.

Chowdhury, A., and C.H. Kirkpatrick. 1994. *Development policy and planning: An introduction to models and techniques*. London: Routledge.

Chung, Y.H., R. Färe, and S. Grosskopf. 1997. Productivity and undesirable outputs: A directional distance function approach. *Journal of Environmental Management* 51: 229–240.

Cicchetti, C.J., A.C. Fisher, and V.K. Smith. 1976. An econometric evaluation of a generalized consumer surplus measure: The Mineral King controversy. *Econometrica* 44: 1259–1276.

Clawson, M. 1959. *Methods of measuring the demand for and value of outdoor recreation, Reprint N°10*. Washington, DC: Resources for the Future Inc.

Coelli, T., G. Hajarhasht, and C.A.K. Lovell. 2008. Econometric estimation of an input distance function in a system of equations. Center for Efficiency and Productivity Analysis, WP01/2008.

Coggins, J.S., and J.R. Swinton. 1996. The price of pollution: A dual approach to valuing SO2 allowances. *Journal of Environmental Economics and Management* 30: 58–72.

Corts, K.S. 1999. Conduct parameters and the measurement of market power. *Journal of Econometrics* 88: 227–250.

Cropper, M.L., L.B. Deck, and K.E. McConnell. 1988. On the choice of functional form for hedonic price functions. *Review of Economics and Statistics* 70: 668–675.

Cross, R., R. Färe, S. Grosskopf, and W.L. Weber. 2013. Valuing vineyards: A Directional distance function approach. *Journal of Wine Economics* 8: 69–82.

Drèze, J., and N. Stern. 1990. Policy reform, shadow prices, and market prices. *Journal of Public Economics* 42: 1–45.

Duygun, M., M. Shaban, R.C. Sickles, and T. Weyman-Jones. 2015. How regulatory capital requirement affects banks' productivity: An Application to emerging economies' banks. *Journal of Productivity Analysis* 44: 237–248.

Farber, S. 1987. The value of coastal wetlands for protection of property against hurricane wind Damage. *Journal of Environmental Economics and Management* 14: 143–151.

Färe, R. 1988. *Fundamentals of production theory*. Lecture Notes in Economics and Mathematical Systems. Berlin: Springer-Verlag.

Färe, R., and S. Grosskopf. 1990. A distance function approach to price efficiency. *Journal of Public Economics* 43: 123–126.

Färe, R., R. Grosskopf, and W.L. Weber. 2006. Shadow prices and pollution costs in the US agriculture. *Ecological Economics* 56: 89–103.

Färe, R., S. Grosskopf, and W.L. Weber. 2001. Shadow prices of Missouri public conservation land. *Public Finance Review* 29: 444–460.

Färe, R., S. Grosskopf, C. Shang, and R.C. Sickles. 2015. Pricing characteristics: An application of Shephard's dual lemma, RISE Working Paper, No. 15-013.

Färe, R., S. Grosskopf, D.-W. Noh, and W. Weber. 2005. Characteristics of a polluting technology: Theory and practice. *Journal of Econometrics* 126: 469–492.

Färe, R., S. Grosskopf, K.C.A. Lovell, and S. Yaisawarng. 1993. Derivation of shadow prices for undesirable outputs: A distance function approach. *Review of Economics and Statistics* 75: 374–380.

Feather, P.M. 1994. Sampling and aggregation issues in random utility model estimation. *American Journal of Agricultural Economics* 76: 772–780.

Fernández, C., G. Koop, and M.F.J. Steel. 2002. Multiple-output production with undesirable outputs: An application to nitrogen surplus in agriculture. *Journal of the American Statistical Association* 97: 432–442.

Formby, J.P., and S. Layson. 1982. Allocative inefficiency and measures of market power. *Atlantic Economic Journal* 10: 67–70.

Førsund, F.R. 2009. Good modelling of bad outputs: Pollution and multiple-output production. *International Review of Environmental and Resource Economics* 3: 1–38.

Freeman III, A.M. 1979. *The benefits of environmental improvement: Theory and practice*. Baltimore, MD: Johns Hopkins University Press.

Freeman III, A.M. 1995. Hedonic Pricing Methods. In *The handbook of environmental economics*, ed. D.W. Bromley, 672–686. Cambridge, MA, USA and Oxford, UK: Blackwell.

Freeman III, A.M. 2003. *The measurement of environmental and resource values: Theory and methods*, 2nd ed. Washington, DC: Resources for the Future.

Getachew, L., and R.C. Sickles. 2007. The policy environment and relative price efficiency of Egyptian private sector manufacturing: 1987/88–1995/96. *Journal of Applied Econometrics* 22: 703–728.

Gollop, F.M., and M.J. Roberts. 1979. Firm interdependence in oligopolistic markets. *Journal of Econometrics* 16: 617–645.

Good, D.H., L. Röller, and R.C. Sickles. 1993. US airline deregulation: Implications for European transport. *Economic Journal* 103: 1028–1041.

Good, D.H., M.I. Nadiri, and R.C. Sickles. 1991. The structure of production, technical change and efficiency in a multinational industry: An application to U.S. airlines. National Bureau of Economic Research Working Paper No. 3939.

Good, D.H., R.C. Sickles, and J. Weiher. 2008. A hedonic price index for airline travel. *Review of Income and Wealth* 54: 438–465.

Graddy, K. 1995. Testing for imperfect competition at the Fulton fish market. *RAND Journal of Economics* 25: 37–57.

Griliches, Z. 1961. Hedonic price indexes for automobiles: An econometric analysis of quality change. In *Price statistics of the federal government*. Washington, DC: U.S. Government Printing Office.

Grosskopf, S., K. Hayes, L.L. Taylor, and W.L. Weber. 1999. Allocative inefficiency and school competition. Proceedings: Ninety-First Annual Conference 1998, National Tax Association, 282–290.

Hammack, J., and G.M. Brown Jr. 1974. *Waterfowl and wetlands: Toward bioeconomic analysis*. Baltimore, MD: Johns Hopkins University Press.

Harberger, A. 1954. Monopoly and resource allocation. *American Economic Review* 44: 77–87.

Harberger, A. 1978. On the use of distributional weights in social cost-benefit analysis. *Journal of Political Economy* 86: 87–120.

Hasannasab, M., D. Margaritis, and C. Staikouras. 2018. The financial crisis and the shadow price of bank capital. Working Paper, University of Auckland.

Hauber, A.B., and G.R. Parsons. 2000. The effect of nesting structure specification on welfare estimation in a random utility model of recreation demand: An application to the demand for recreational fishing. *American Journal of Agricultural Economics* 82: 501–514.

Hausman, J. 2012. Contingent valuation: From dubious to hopeless. *Journal of Economic Perspectives* 26: 43–56.

Hellerstein, D.M. 1991. Using count data models in travel cost analysis with aggregate data. *American Journal of Agricultural Economics* 73: 860–866.

Hellerstein, D.M. 1992. The treatment of nonparticipants in travel cost analysis and other demand models. *Water Resources Research* 28: 1999–2004.

Hellerstein, D.M., and R. Mendelsohn. 1993. A theoretical foundation for count data models. *American Journal of Agricultural Economics* 75: 604–611.

Hetemäki, L. 1996. Essays on the impact of pollution control on a firm: A distance function approach. Research Papers, vol. 609. The Finnish Forest Research Institute, Helsinki.

Hicks, J. 1935. Annual survey of economic theory: The theory of monopoly. *Econometrica* 3: 1–20.

Irwin, E.G., and N.E. Bockstael. 2001. The problem of identifying land use spillovers: Measuring the effects of open space on residential property values. *American Journal of Agricultural Economics* 83: 698–704.

Iwata, G. 1974. Measurement of conjectural variations in oligopoly. *Econometrica* 42: 947–966.

Jeon, B.M., and R.C. Sickles. 2004. The role of environmental factors in growth accounting. *Journal of Applied Econometrics* 19: 567–591.

Karakaplan, M.U., and L. Kutlu. 2019. Estimating market power using a composed error model. *Scottish Journal of Political Economy* 66: 489–510.

Kling, C.L., D.J. Phaneuf, and J. Zhao. 2012. From Exxon to BP: Has some number become better than no number? *Journal of Economic Perspectives* 26: 3–26.

Koetter, M., J.W. Kolari, and L. Spierdijk. 2012. Enjoying the quiet life under deregulation? *Evidence from Adjusted Lerner Indices for US Banks, Review of Economics and Statistics* 94: 462–480.

Kumbhakar, S.C., S. Baardsen, and G. Lien. 2012. A new method for estimating market power with an application to Norwegian sawmilling. *Review of Industrial Organization* 40: 109–129.

Kuminoff, N.V., C.F. Parmeter, and J.C. Pope. 2010. Which hedonic models can we trust to recover the marginal willingness to pay for environmental amenities? *Journal of Environmental Economics and Management* 60: 145–160.

Kutlu, L. 2017. A conduct parameter model of price discrimination. *Scottish Journal of Political Economy* 64: 530–536.

Kutlu, L., and R.C. Sickles. 2012. Estimation of market power in the presence of firm level inefficiencies. *Journal of Econometrics* 168: 141–155.

Kutlu, L., and R.C. Sickles. 2017. Measuring market power when firms price discriminate. *Empirical Economics* 53: 287–305.

Kutlu, L., and R. Wang. 2018. Estimation of cost efficiency without cost data. *Journal of Productivity Analysis* 49: 137–151.

Lau, L.J. 1982. On identifying the degree of competitiveness from industry price and output Data. *Economics Letters* 10: 93–99.

Lau, L.J., and P.A. Yotopoulos. 1971. A test for relative efficiency and application to Indian agriculture. *American Economic Review* 61: 94–109.

Lee, J.D., J.B. Park, and T.Y. Kim. 2002. Estimation of the shadow prices of pollutants with production/environment efficiency taken into account: A nonparametric directional distance function approach. *Journal of Environmental Management* 64: 365–375.

Lerner, A.P. 1934. The concept of monopoly and measurement of monopoly power. *Review of Economic Studies* 1: 157–175.

Lindhjem, H., and S. Navrud. 2011. Using Internet in stated preference surveys: A review and comparison of survey modes. *International Review of Environmental and Resource Economics* 5: 309–351.

Little, I., and J. Mirrlees. 1969. *Manual of industrial project analysis in developing countries*, vol. 2. Paris: Organisation for Economic Cooperation and Development.

Little, I., and J. Mirrlees. 1974. *Project appraisal and planning for developing countries*. London: Heinemann.

Lovell, C.A.K., and R.C. Sickles. 1983. Testing efficiency hypotheses in joint production: A parametric approach. *Review of Economics and Statistics* 65 (1): 51–58.

MacMillan, D., N. Hanley, and N. Leinhoop. 2006. Contingent valuation: Environmental polling or preference engine? *Ecological Economics* 60: 299–307.

Malikov, E., R. Bokusheva, and S.C. Kumbhakar. 2016. A hedonic output index based approach to modeling polluting technologies. *Empirical Economics* 54 (1): 287–308.

Maudos, J., and J.F. Guevara. 2007. The cost of market power in banking: Social welfare loss vs. cost inefficiency. *Journal of Banking & Finance* 31: 2103–2125.

Michener, R., and C. Tighe. 1992. A Poisson regression model of highway fatalities. *American Economic Review* 82: 452–456.

Mitchell, R.C., and R.T. Carson. 1989. *Using surveys to value public goods: The contingent valuation method*. Washington, DC: Resources for the Future.

Munro, A. 2005. Household willingness to pay equals individual willingness to pay if and only if the household income pools. *Economics Letters* 88: 227–230.

Murty, S., R.R. Russell, and S.B. Levkoff. 2012. On modeling pollution-generating technologies. *Journal of Environmental Economics and Management* 64: 117–135.

O'Donnell, C.J., G.R. Griffith, J.J. Nightingale, and R.R. Piggott. 2007. Testing for market power in the Australian grains and oilseeds industries. *Agribusiness* 23: 349–376.

Olson, M.K. 2004. Are novel drugs more risky for patients than less novel drugs? *Journal of Health Economics* 23: 1135–1158.

Orea, L., and J. Steinbuks. 2012. Estimating market power in homogenous product markets using a composed error model: Application to the California electricity market. University of Cambridge, Faculty of Economics.

Overseas Development Administration. 1988. *Appraisal of projects in developing countries*. London: Her Majesty's Stationery Office.

Parsons, K.J. 2017. Travel cost models, A primer on nonmarket valuation. In *The economics of non-market goods and resources*, vol. 13, 187–233. Dordrecht, The Netherlands: Springer.

Parsons, G.R., and M.J. Kealy. 1992. Randomly drawn opportunity sets in a random utility model of lake recreation. *Land Economics* 68: 93–106.

Perloff, J.M., and E.Z. Shen. 2012. Collinearity in linear structural models of market power. *Review of Industrial Organization* 2: 131–138.

Perloff, J.M., L.S. Karp, and A. Golan. 2007. *Estimating market power and strategies*. Cambridge: Cambridge University Press.

Phaneuf, D.J., and T. Requate. 2016. *A course in environmental economics: Theory, policy, and practice*. New York: Cambridge University Press.

Pindyck, R.S. 1985. The measurement of monopoly power in dynamic markets. *Journal of Law and Economics* 28: 193–222.

Plane, P. 1999. Privatization, technical efficiency and welfare consequences: The case of the Côte d'Ivoire Electricity Company (CIE). *World Development* 27 (2): 343–360.

Poe, G.L., and R.C. Bishop. 1999. Valuing the incremental benefits of groundwater protection when exposure levels are known. *Environmental & Resource Economics* 13: 347–373.

Porter, R. 1983. A study of cartel stability: The joint executive committee 1980–1986. *Bell Journal of Economics* 14: 301–314.

Primont, D.F., and M.F. Kokoski. 1990. Comparing prices across cities: A hedonic approach. BLS Working Papers #204.

Puller, S.L. 2007. Pricing and firm conduct in California's deregulated electricity market. *Review of Economics and Statistics* 89: 75–87.

Puller, S.L. 2009. Estimation of competitive conduct when firms are efficiently colluding: Addressing the Corts critique. *Applied Economics Letters* 16: 1497–1500.

Quiggin, J. 1998. Individual and household willingness to pay for public goods. *American Journal of Agricultural Economics* 80: 58–63.

Quigley, J.M. 1982. Nonlinear budget constraints and consumer demand: An application to public programs for residential housing. *Journal of Urban Economics* 12: 177–201.

Randall, A., B. Ives, and C. Eastman. 1974. Bidding games for evaluation of aesthetic environmental improvements. *Journal of Environmental Economics and Management* 1: 132–149.

Ray, A. 1984. *Cost-Benefit Analysis*. Baltimore, MD: Johns Hopkins University Press.

Rhoades, S.A. 1982. Welfare loss, redistribution effect, and restriction in output due to monopoly in banking. *Journal of Monetary Economics* 9: 375–387.

Ridker, R.G., and J.A. Henning. 1967. The determinants of residential property values with special reference to air pollution. *Review of Economics and Statistics* 49: 246–257.

Roehrig, C.S. 1988. Conditions for identification in nonparametric and parametric models. *Econometrica* 56: 433–447.

Röller, L.H., and R.C. Sickles. 2000. Capacity and product market competition: Measuring market power in a 'puppy-dog' industry. *International Journal of Industrial Organization* 18: 845–865.

Rose, N.L. 1990. Profitability and product quality: Economic determinants of airline safety performance. *Journal of Political Economy* 98: 944–964.

Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy* 82: 34–55.

Rosenthal, R. 1980. A model in which an increase in the number of sellers leads to a higher price. *Econometrica* 48: 1575–1599.

Schwartz, A.E., and B.P. Scafidi. 2000. Quality adjusted price indices for four year colleges, Mimeo, New York University and Georgia State University.

Shaw, D. 1988. On-site samples' regression: Problems of non-negative integers, truncation, and endogenous stratification. *Journal of Econometrics* 37: 211–223.

Shephard, R.W. 1953. *Cost and production functions*. Princeton: Princeton University Press.

Shephard, R.W. 1970. *Theory of cost and production functions*. Princeton: Princeton University Press.

Sickles, R.C., and M.L. Streitwieser. 1998. An analysis of technology, productivity, and regulatory distortion in the interstate natural gas transmission industry: 1977–1985. *Journal of Applied Econometrics* 13: 377–395.

Sickles, R.C., D. Good, and R.L. Johnson. 1986. Allocative distortions and the regulatory transition of the U.S. airline industry. *Journal of Econometrics* 33: 143–163.

Spiller, P.T., and E. Favaro. 1984. The effects of entry regulation or oligopolistic interaction: The Uruguayan banking sector. *RAND Journal of Economics* 15: 244–254.

Squire, L., and H. van der Tak. 1975. *Economic analysis of projects*. Baltimore, MD: Johns Hopkins University Press.

Stiglitz, J. 1989. Imperfect information in the product market. In *The handbook of industrial organization*, vol. 1, ed. Richard Schmalensee and Robert D. Willig, 769–847. Amsterdam: North-Holland.

Stole, L.A. 2007. Price discrimination and competition. In *The handbook of industrial organization*, vol. 3, ed. M. Armstrong and R. Porter, 2221–2299. Amsterdam: North-Holland.

Swinton, J.R. 1998. At what cost do we reduce pollution? Shadow prices of SO2 emissions. *Energy Journal* 19: 63–83.

Taylor, L.O. 2017. Hedonics, A primer on nonmarket valuation. In *The economics of non-market goods and resources*, vol. 13, 235–292. Dordrecht, The Netherlands: Springer.

Timmins, C., and J. Murdock. 2007. A revealed preference approach to the measurement of congestion in travel cost models. *Journal of Environmental Economics and Management* 53: 230–249.

Train, K.E. 1998. Recreation demand models with taste differences over people. *Land Economics* 74: 230–239.

Trice, A.H., and S.E. Wood. 1958. Measurement of recreation benefits. *Land Economics* 34: 195–207.

Triplett, J.E. 1984. Measuring technological change with characteristics-space technique. BLS Working Papers #141.

Triplett, J.E. 2004. *Handbook on hedonic indexes and quality adjustments in price indexes: Special application to information technology products*. Washington, DC: Brookings Institution.

Tsionas, E., S.C. Kumbhakar, and E. Malikov. 2015. Estimation of input distance functions: A system approach, Munich Personal RePEc Archive.

UNIDO. 1972. *Guidelines for project evaluation*. New York: United Nations Industrial Development Organization.

Vardanyan, M., and D.-W. Noh. 2006. Approximating pollution abatement costs via alternative specifications of a multi-output production technology: A case

of the U.S. electric utility industry. *Journal of Environmental Management* 80: 177–190.

Waugh, F.V. 1928. Quality factors influencing vegetable prices. *Journal of Farm Economics* 10: 185–196.

Weiher, J.C., R.C. Sickles, and J. Perloff. 2002. Market power in the US airline industry. In *Measuring market power*, ed. D.J. Slottje, 309–323. Amsterdam: North-Holland.

Winkelmann, R. 2003. *Econometric analysis of count data*, 4th ed. Berlin: Springer-Verlag.

Aggregation of Individual Efficiency Measures and Productivity Indices

Andreas Mayer and Valentin Zelenyuk

1 Introduction

While analyzing the efficiency or productivity of organizations (firms, banks, hospitals, or their departments, or industries or countries, etc.), it is usually of interest to consider not only the efficiency of each individual but also (and often more importantly) an aggregate measure summarizing the efficiency of a group. To do so, researchers often used simple averages of individual measures or indices. Such an approach, however, ignores the economic weight of each organization whose efficiency or productivity scores are averaged, and so such simple or equally weighted averages can severely misrepresent the situation.

Consider, for example, a case when an industry has one or very few large firms and many small firms. In fact, many industries in practice display this scenario. Suppose that the large firm(s) happens to be very inefficient for various reasons: e.g., due to being almost monopolies, due to better political or financial connections or status ('too big to fail' phenomenon), etc.

A. Mayer
Centre for Efficiency and Productivity Analysis,
The University of Queensland, Brisbane, QLD, Australia

V. Zelenyuk (✉)
School of Economics and Centre for Efficiency and Productivity Analysis,
The University of Queensland, Brisbane, QLD, Australia
e-mail: v.zelenyuk@uq.edu.au

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_14

On the other hand, suppose the hundreds of small firms cannot afford to be inefficient and so are highly competing for 'their place under the sun' so to speak, and thus attain very high efficiency scores. If one applies the simple average to aggregate the individual efficiency scores in such an industry, then this aggregate measure of efficiency will give a very high aggregate efficiency score—because the simple average ignores the economic weight, assigning equal weight to each firm. Indeed, even if the majority of the small firms take only say 1% of the industry share, while each being say 100% efficient, and the rest of the industry is dominated by the large say 50% efficient firm(s), the simple average will give the aggregate score close to 100% efficiency, while the reality of such an industry is that it operates at a very inefficient (about 50%) level. Using some economic weighting of the efficiency scores, entering the averaging therefore appears to be very useful, if not critical for analysis and policy implications.

Not surprisingly, such questions of measuring aggregate or group efficiency have been explored extensively in the last few decades. It goes back to at least Farrell (1957), who proposed what he dubbed the 'structural efficiency of an industry'—his idea was to use the observed output shares of firms in a group (in a single output case) to weight the individual efficiency scores (input-oriented and estimated under the assumption of constant returns to scale) of these firms. Farrell gave intuition for such an aggregation scheme but did not derive it from any assumptions or reasoning and so it remained an ad hoc aggregate measure.

Farrell's idea of an industry efficiency measure was then explored in many works. One of the earliest works on this was due to Førsund and Hjalmarsson (1979), who appears to be the first who introduced the concept of the efficiency of the average decision-making unit. This idea was further elaborated by the very insightful work of Li and Ng (1995), who appear to be the first to derive a scheme of aggregation using shadow prices, where the aggregate revenue efficiency is decomposed into aggregate technical efficiency and aggregate allocative efficiency. On the purely theoretical level, the aggregation question was also explored by Blackorby and Russell (1999) who discovered several impossibility results for a general efficiency aggregation problem. Various ideas of aggregate efficiency were also critically discussed in an influential article by Ylvinger (2000), who pointed out various problems with the existing approaches and suggested aggregation using weights derived from data envelopment analysis (DEA).[1]

[1]The name DEA was coined by Charnes et al. (1978), who generalized the approach of Farrell (1957) to the multi-output case, refined it from the mathematical programming perspective and immensely

More recently, Färe and Zelenyuk (2003), synthesized the existing results on aggregation, and in particular modified the approach of Li and Ng (1995) by modifying the important (yet forgotten at that time) results from Koopmans (1957) and proposed a new approach for aggregation of Farrell-type individual efficiency scores, where the aggregation function and the weights are derived from an economic theory perspective under certain assumptions. The weights that were derived turned out to be observed revenue shares (for output-oriented measurement) or observed cost shares (for the input orientation). In the special case of single output, they become the observed output shares, i.e., the same weights as in Farrell's 'structural efficiency of an industry' (although note that his ad hoc measure was aggregated input-oriented efficiency scores).

The strategy of Färe and Zelenyuk (2003) for solving the aggregation problem was then used to derive similar aggregation results in other related contexts. For example, Färe et al. (2004) applied it for the input (cost) orientation context and Färe and Zelenyuk (2005) used it further to suggest the aggregation weights for cases when one wants to use the geometric aggregating function, which they derived using an alternative approach, based on solving functional equations. Furthermore, Zelenyuk (2006) used a similar strategy, extended to the intertemporal context, to derive aggregation results for the Malmquist productivity index (MPI, see Caves et al. (1982)) and its decompositions. A similar solution strategy was also applied by Zelenyuk (2011) to resolve the problem of aggregating individual growth rates and its sources in the Solow-type growth accounting framework. Meanwhile, the problem of aggregating individual scale elasticities was considered and solved by Färe and Zelenyuk (2012), while Zelenyuk (2015) resolved the problem of aggregating individual scale efficiencies.

It is important to note that the aggregation approach from Färe and Zelenyuk (2003) and all those mentioned in the paragraph above required a key assumption that the input endowment across firms in the group is fixed and cannot be reallocated between the individual organizations in the group. This assumption was later relaxed by Nesterenko and Zelenyuk (2007) who proposed a more general approach, embracing that of Färe and Zelenyuk (2003) as a special case.[2] Specifically, they proposed new output-oriented group efficiency measures that allow for inputs to be reallocated between decision-making units within the group. This relaxation is particularly important for contexts where reallocation of inputs is possible, e.g., when

popularized it in the business and management science research. While this is perhaps the most popular approach to estimate the Farrell efficiency, other methods can be used as well, and most of the discussion here is general, for any suitable estimators, unless specified otherwise.

[2]Also see Färe et al. (2008) for aggregation of efficiencies based on directional distance functions.

the analyzed organizations are branches within a larger unity (e.g., departments as part of a firm), when two or more organizations are merging, or when countries are uniting into an economic union. For instance, consider an organization (bank, hospital, etc.) with multiple sub-units, which can move its staff and other resources (capital, materials, energy, etc.) between these sub-units. In this case, treating the inputs as fixed when measuring aggregate efficiency of this organization without accounting for the possibility or reallocation of inputs seems inadequate. More recently, ten Raa (2011) arrived at a similar result, deriving it from a different angle and showing a new relationship to concepts from the field of industrial organization.

The ideas from Nesterenko and Zelenyuk (2007) were then elaborated by Mayer and Zelenyuk (2014a) to generalize the aggregation approach developed by Zelenyuk (2006) for the Malmquist productivity index of Caves et al. (1982).[3]

While MPI remains the most popular approach for measuring productivity changes, more attention has been given recently to an alternative approach—the Hicks-Moorsteen Productivity Index (HMPI), introduced by Diewert (1992) and Bjurek (1996). The HMPI has some appealing theoretical and practical properties; e.g., it always has a total factor productivity (TFP) interpretation. That is, it has the interpretation of measuring a change in aggregate output relative to the change in aggregate input, an intuitive notion of productivity for a multi-output economic unit (see Epure et al. 2011). Like the MPI, the HMPI uses the Shephard (1953, 1970) distance functions to calculate the resulting productivity change for an individual organization, without including price information. Aggregation for HMPI was outlined in the working paper of Mayer and Zelenyuk (2014b), and here, we present a brief and refined version of it.

In a nutshell, the rest of the paper aims to fulfill three goals: The first goal is to summarize some of the key existing results on aggregate efficiency measures and aggregate productivity indices. Reaching this goal will outline the foundation and the building blocks for the second, and as important, goal—to outline the results for aggregation of HMPI. The third goal is more modest, yet still important—it is to outline some insights on ongoing and future directions of research in this area.

In the next section, we present the individual efficiency and productivity measures. Section 3 presents the key aggregation results for the group efficiency measures. Section 4 presents the main results, while Sect. 5 considers

[3]Also note that this is an alternative approach in the theory of index number aggregation (e.g., see Diewert (1983, 1985) and references therein). For applications of these type of indexes, see a review by Badunenko et al. (2017).

practical issues about estimation and deriving price-independent weights. Section 6 concludes. For a related (shorter and simplified) discussion, also see Chapter 5 of Sickles and Zelenyuk (2019).

2 Individual Efficiency and Productivity Measures

Let us start by considering individual efficiency measures for a group of K organizations. In different contexts, this group could be a country consisting of regions, or an industry consisting of firms, banks, hospitals, or a firm consisting of departments, etc. Suppose an organization $k(k = 1, \ldots, K)$ uses vector $x^k = \left(x_1^k, \ldots, x_N^k\right)' \in \mathbb{R}_+^N$ of N inputs to produce a vector $y^k = \left(y_1^k, \ldots, y_M^k\right)' \in \mathbb{R}_+^M$ of M outputs. For a given organization k and time period τ, we assume the organization's technology can be expressed as the *technology set* T_τ^k:

$$T_\tau^k = \left\{(x, y) \in \mathbb{R}_+^N \times \mathbb{R}_+^M : \text{organization } k \text{ can produce } y \text{ from } x \text{ in period } \tau\right\}. \quad (1)$$

We will also use two alternative and equivalent characterizations via the *input requirement correspondence*, $L_\tau^k : \mathbb{R}_+^M \to 2^{\mathbb{R}_+^N}$:

$$L_\tau^k\left(y^k\right) = \left\{x^k \in \mathbb{R}_+^N : \left(x^k, y^k\right) \in T_\tau^k\right\}, \quad y^k \in \mathbb{R}_+^M. \quad (2)$$

and the *output correspondence*, $P_\tau^k : \mathbb{R}_+^N \to 2^{\mathbb{R}_+^M}$:

$$P_\tau^k\left(x^k\right) = \left\{y^k \in \mathbb{R}_+^M : \left(x^k, y^k\right) \in T_\tau^k\right\}, \quad x^k \in \mathbb{R}_+^N. \quad (3)$$

We accept the standard regularity axioms of production theory (see Färe and Primont (1995) and Sickles and Zelenyuk (2019)) for more detailed discussions. Specifically (for all $k = 1, \ldots, K$ and for all τ), we assume:

Axiom 1 The technology set T_τ^k is closed.

Axiom 2 The output correspondence $P_\tau^k\left(x^k\right)$ is bounded for all $x^k \in \mathbb{R}_+^N$.

Axiom 3 There is no 'free lunch,' i.e., nothing cannot produce something, i.e., $\left(0_N, y^k\right) \notin T_\tau^k$, for all $y^k \geq 0_M$ (i.e., $y_m^k \geq 0$ for $m = 1, \ldots, M, y^k \neq 0_M$).

Axiom 4 It is possible to produce nothing, i.e., $0_M \in P_\tau^k\left(x^k\right)$, for all $x^k \in \mathbb{R}_+^N$.[4]

Axiom 5 Outputs and inputs are freely (strongly) disposable, i.e., $\left(x^0, y^0\right) \in T_\tau^k \Rightarrow (x, y) \in T_\tau^k$, for all $y \leqq y^0$, for all $x \geqq x^0$, $y \geq 0$.

Axiom 6 Output correspondences $P_\tau^k\left(x^k\right)$ are convex, for all $x^k \in \mathbb{R}_+^N$.

Axiom 7[5] Input requirement correspondences $L_\tau^k\left(y^k\right)$ are convex, for all $y^k \in \mathbb{R}_+^M$.

2.1 Individual Efficiency Measures

Following the literature (e.g. see Sickles and Zelenyuk (2019) and detailed references there), the *input-* and *output-oriented Farrell-type technical efficiency* measures for an individual organization k in period τ are defined, respectively, as:

$$\mathrm{ITE}_\tau^k = \mathrm{ITE}_\tau^k\left(y^k, x^k\right) = \inf\left\{\lambda \in \mathbb{R}_{++} : \lambda x^k \in L_\tau^k\left(y^k\right)\right\}, \quad \left(x^k, y^k\right) \in \mathbb{R}_+^N \times \mathbb{R}_+^M. \tag{4}$$

and

$$\mathrm{OTE}_\tau^k = \mathrm{OTE}_\tau^k(x^k, y^k) = \sup\left\{\theta \in \mathbb{R}_{++} : \theta y^k \in P_\tau^k\left(x^k\right)\right\}, \quad \left(x^k, y^k\right) \in \mathbb{R}_+^N \times \mathbb{R}_+^M. \tag{5}$$

This allows us to measure the efficiency of a organization k without needing to take prices into account.[6] Likewise the dual characterization of technology, the *cost* and *revenue functions* are defined, respectively, as:

[4]Note that Axioms 3 and 4 together imply that $(0, 0) \in T^k$ for all k.

[5]We require Axioms 6 and 7 for duality results to hold. For our theoretical results, we do not require convexity of T^k, but when discussing practical methods of estimation we then introduce this stronger assumption.

[6]Note that some authors define the output oriented Farrell-type technical efficiency measures as the reciprocal of (5), which range in value between 0 and 1. The derivations following could be rewritten with this definition, but the definition we have given appears to be the more common definitions in the literature, particularly in previous works on aggregation of efficiency, so we continue with it.

$$C_\tau^k\left(y^k,w\right)=\inf_x\left\{wx:x\in L_\tau^k\left(y^k\right)\right\},\ y^k\in\mathbb{R}_+^M,\ w\in\mathbb{R}_{++}^N, \tag{6}$$

and

$$R_\tau^k\left(x^k,p\right)=\sup_y\left\{py:y\in P_\tau^k\left(x^k\right)\right\},\ x^k\in\mathbb{R}_+^N,\ p\in\mathbb{R}_{++}^M, \tag{7}$$

given an input price row-vector $w=(w_1,\ldots,w_N)\in\mathbb{R}_{++}^N$ corresponding to the N inputs and output price row-vector $p=(p_1,\ldots,p_M)\in\mathbb{R}_{++}^M$ corresponding to the M outputs. Throughout, we make the assumption that w and p are identical across every organization, which is necessary to derive the aggregation results discussed in Sect. 3. The *cost* and *revenue efficiencies* of organization k at period τ are, respectively:

$$\mathrm{CE}_\tau^k=\mathrm{CE}_\tau^k\left(y^k,x^k,w\right)=\frac{C_\tau^k\left(y^k,w\right)}{wx^k},\quad \text{for}\, wx^k\neq 0,\ \left(x^k,y^k\right)\in\mathbb{R}_+^N\times\mathbb{R}_+^M, \tag{8}$$

and

$$\mathrm{RE}_\tau^k=\mathrm{RE}_\tau^k\left(x^k,y^k,p\right)=\frac{R_\tau^k\left(x^k,p\right)}{py^k},\quad \text{for}\, py^k\neq 0,\ \left(x^k,y^k\right)\in\mathbb{R}_+^N\times\mathbb{R}_+^M. \tag{9}$$

Note that in all cases, for $\left(x^k,y^k\right)\in T_\tau^k$, we have $\mathrm{CE}^k\left(y^k,x^k,w\right)\leq\mathrm{ITE}^k\left(y^k,x^k\right)$ and $\mathrm{RE}^k\left(x^k,y^k,p\right)\geq\mathrm{OTE}^k\left(x^k,y^k\right)$, and these inequalities can be closed by introducing the multiplicative residuals referred to as *input-* and *output-oriented allocative efficiency*, respectively, defined as:

$$\mathrm{IAE}_\tau^k=\mathrm{IAE}_\tau^k\left(y^k,x^k,w\right)=\frac{\mathrm{CE}_\tau^k\left(y^k,x^k,w\right)}{\mathrm{ITE}_\tau^k\left(y^k,x^k\right)},\quad \left(x^k,y^k\right)\in\mathbb{R}_+^N\times\mathbb{R}_+^M, \tag{10}$$

and

$$\mathrm{OAE}_\tau^k=\mathrm{OAE}_\tau^k\left(x^k,y^k,p\right)=\frac{\mathrm{RE}_\tau^k\left(x^k,y^k,p\right)}{\mathrm{OTE}_\tau^k\left(x^k,y^k\right)},\quad \left(x^k,y^k\right)\in\mathbb{R}_+^N\times\mathbb{R}_+^M. \tag{11}$$

So, one gets the following decompositions for any period τ and any organization k,

$$\begin{aligned} \mathrm{CE}_{\tau}^{k}\left(y^{k}, x^{k}, w\right) &= \mathrm{ITE}_{\tau}^{k}\left(y^{k}, x^{k}\right) \times \mathrm{IAE}_{\tau}^{k}\left(y^{k}, x^{k}, w\right), \\ &\text{for all } \left(x^{k}, y^{k}\right) \in \mathbb{R}_{+}^{N+M},\ wx^{k} \neq 0, \end{aligned} \tag{12}$$

and

$$\begin{aligned} \mathrm{RE}_{\tau}^{k}\left(x^{k}, y^{k}, p\right) &= \mathrm{OTE}_{\tau}^{k}\left(x^{k}, y^{k}\right) \times \mathrm{OAE}_{\tau}^{k}\left(x^{k}, y^{k}, p\right), \\ &\text{for all } \left(x^{k}, y^{k}\right) \in \mathbb{R}_{+}^{N+M},\ py^{k} \neq 0, \end{aligned} \tag{13}$$

Later on, we will want aggregate analogues of such decompositions to hold as well. It is also worth noting that all these efficiency measures can be deduced as special cases or components of the general profit efficiency measure in Färe et al. (2019) and, in particular, related to the profit-optimization paradigm.

It is also worth noting that there are extreme cases when the functions discussed in these sections attain zero or ∞, which is problematic when using these functions as building blocks of productivity indices. For this reason, from now on we explicitly assume and limit our discussion to only cases when these functions attain finite and strictly positive values. Also, see related discussion on regularizations in Färe et al. (2019).

2.2 Individual Productivity Indices

Following Diewert (1992) and Bjurek (1996), the *Hicks-Moorsteen productivity index* (HMPI) for measuring the productivity change from period s to t (for organization k) can be defined as:

$$\mathrm{HM}_{st}^{k} = \mathrm{HM}_{st}^{k}\left(y_{s}^{k}, y_{t}^{k}, x_{s}^{k}, x_{t}^{k}\right) = \left[\left(\frac{\frac{\mathrm{OTE}_{s}^{k}\left(x_{s}^{k}, y_{t}^{k}\right)}{\mathrm{OTE}_{s}^{k}\left(x_{s}^{k}, y_{s}^{k}\right)}}{\frac{\mathrm{ITE}_{s}^{k}\left(y_{s}^{k}, x_{t}^{k}\right)}{\mathrm{ITE}_{s}^{k}\left(y_{s}^{k}, x_{s}^{k}\right)}} \times \frac{\frac{\mathrm{OTE}_{t}^{k}\left(x_{t}^{k}, y_{t}^{k}\right)}{\mathrm{OTE}_{t}^{k}\left(x_{t}^{k}, y_{s}^{k}\right)}}{\frac{\mathrm{ITE}_{t}^{k}\left(y_{t}^{k}, x_{t}^{k}\right)}{\mathrm{ITE}_{t}^{k}\left(y_{t}^{k}, x_{s}^{k}\right)}}\right)^{-1}\right]^{1/2}. \tag{14}$$

Note that there is now a time subscript for inputs and outputs in the components representing input- and output-oriented efficiency, unlike in the previous section. In particular, note that for the output-oriented measures, inputs

are held constant at the same period as the technology (while outputs are varied), and for the input-oriented measures, outputs are held constant at the same period as the technology (while inputs are varied). Also note that, of course, one can get rid of the reciprocal by flipping all the ratios, yet here and everywhere below we decided not to do so to remind and emphasize that we are using Farrell-type technical measures rather than their reciprocals (or Shephard's distance functions), as is typically done in the literature.

In light of the duality between the revenue or cost efficiency measures with the technical efficiency measures, one can define a dual Hicks-Moorsteen productivity index, analogous to the original HMPI but with price information, i.e.,

$$\begin{aligned}\mathrm{PHM}_{st}^{k} &= \mathrm{PHM}_{st}^{k}\left(y_s^k, y_t^k, x_s^k, x_t^k, p_s, p_t, w_s, w_t\right) \\ &= \left[\left(\frac{\frac{\mathrm{RE}_s^k\left(x_s^k, y_t^k, p_t\right)}{\mathrm{RE}_s^k\left(x_s^k, y_s^k, p_s\right)}}{\frac{\mathrm{CE}_s^k\left(y_s^k, x_t^k, w_t\right)}{\mathrm{CE}_s^k\left(y_s^k, x_s^k, w_s\right)}} \times \frac{\frac{\mathrm{RE}_t^k\left(x_t^k, y_t^k, p_t\right)}{\mathrm{RE}_t^k\left(x_t^k, y_s^k, p_s\right)}}{\frac{\mathrm{CE}_t^k\left(y_t^k, x_t^k, w_t\right)}{\mathrm{CE}_t^k\left(y_t^k, x_s^k, w_s\right)}}\right)^{-1}\right]^{1/2},\end{aligned} \tag{15}$$

and we will refer to this measure as the *profitability Hicks-Moorsteen productivity index* from periods s to t for organization k. This name seems justified because for any $\tau, j \in \{s, t\}$ the measure can alternatively be represented in terms of profitability components (optimal and observed) of the following form:

$$\frac{\mathrm{RE}_\tau^k\left(x_\tau^k, y_j^k, p_j\right)}{\mathrm{CE}_\tau^k\left(y_\tau^k, x_j^k, w_j\right)} = \frac{R_\tau^k\left(x_\tau^k, p_j\right)/C_\tau^k\left(y_\tau^k, w_j\right)}{p_j y_j^k / w_j x_j^k}. \tag{16}$$

As before, outputs (inputs) and technology are kept in the same periods for input- (output-) oriented measures, while the input (output) prices are permitted to vary with the inputs (outputs) for the input- (output-) oriented measures. In other words, by measuring cost/revenue w.r.t. a given orientation, the organization is treated as making choices regarding that factor (inputs or outputs), taking as given the amount of the other factor and the prices of the first factor in that period.

Before going further, let us pause here and consider the meaning of (15). Although rarely used in the literature, (15) may actually be seen as superior to the more frequently used primal HMPI (14), because it also takes into account such important economic information as prices (and potential

inefficiency with respect to them). Meanwhile, the productivity change between periods s and t measured by the primal HMPI ignores the price information completely (both w.r.t. period s technology (the initial fraction) and w.r.t. period t technology).

In light of the duality results we mentioned in the previous section, substituting (12) and (13) into (15) and arranging terms to isolate (14), we immediately get the following useful decomposition

$$\mathrm{PHM}^k_{st} = \mathrm{HM}^k_{st} \times \mathrm{AHM}^k_{st}, \tag{17}$$

where the last term is a remainder, defined as

$$\begin{aligned}\mathrm{AHM}^k_{st} &= \mathrm{AHM}^k_{st}\left(y^k_s, y^k_t, x^k_s, x^k_t, p_s, p_t, w_s, w_t\right) \\ &\equiv \left[\left(\frac{\frac{\mathrm{OAE}^k_s\left(x^k_s, y^k_t, p_t\right)}{\mathrm{OAE}^k_s\left(x^k_s, y^k_s, p_s\right)}}{\frac{\mathrm{IAE}^k_s\left(y^k_s, x^k_t, w_t\right)}{\mathrm{IAE}^k_s\left(y^k_s, x^k_s, w_s\right)}} \times \frac{\frac{\mathrm{OAE}^k_t\left(x^k_t, y^k_t, p_t\right)}{\mathrm{OAE}^k_t\left(x^k_t, y^k_s, p_s\right)}}{\frac{\mathrm{IAE}^k_t\left(y^k_t, x^k_t, w_t\right)}{\mathrm{IAE}^k_t\left(y^k_t, x^k_s, w_s\right)}}\right)^{-1}\right]^{1/2},\end{aligned} \tag{18}$$

which can be called the *allocative Hicks-Moorsteen productivity index* from periods s to t for organization k. This decomposition holds for any input–output prices combination, in any periods s and t, and for all k.

In words, (17) suggests that our profitability index can be decomposed into a measure of productivity change with and without accounting for prices: The primal HMPI considers productivity changes using inputs and outputs but not prices, whereas the allocative HMPI considers productivity changes both due to changes in the allocation of inputs/outputs and in addition to changes between periods for each price. Given the value of this decomposition, an analogous decomposition at the aggregate level is desirable, which is the topic we discuss next. For more of theoretical and practical details of these and other indexes, see Chapter 4 and Chapter 7 of Sickles and Zelenyuk (2019).[7]

[7]It is worth noting that there is also an additive form of the Hicks-Moorsteen productivity index, often referred to as the Luenberger–Hicks-Moorsteen productivity indicator (e.g., see Briec and Kerstens 2004). The aggregation theory for such an index can also be developed in a similar fashion based on aggregation results for directional distance functions from Färe et al. (2008).

3 Aggregate Efficiency Measures

We now consider aggregate efficiency measures which measure the efficiency of a group of organizations, which will be useful in constructing our aggregate HMPI measures in Sect. 4. For ease of notation, we present results for aggregating all organizations in a group. (One can extend it to aggregation into subgroups of organizations, which can then be consistently aggregated into larger groups, similarly to Simar and Zelenyuk (2007), at a cost of more complex notation and some additional derivations.) In the interests of space, we also only present results for input orientation for this section; the formulae and derivations for output orientation are analogous and readers interested in more details can find them in Färe and Zelenyuk (2003), Zelenyuk (2006), Nesterenko and Zelenyuk (2007), Mayer and Zelenyuk (2014a). Also see ten Raa (2011) for related derivations and interesting discussions in the context of industrial organization. While we focus on input and output orientations, readers interested in applying aggregation to directional distance functions are referred to Briec et al. (2003) and Färe et al. (2008) and the references cited therein. Also, for a related (and more introductory) discussion, also see Chapter 5 of Sickles and Zelenyuk (2019).

3.1 Aggregate Efficiency Measures with Restrictions on Reallocation

We denote the input and output allocations among organizations within the group at a given period τ as $X_\tau = \left(x_\tau^1, \ldots, x_\tau^K\right)$ and $Y_\tau = \left(y_\tau^1, \ldots, y_\tau^K\right)$, and the sum of these over all organizations in the group as $\overline{X}_\tau = \sum_{k=1}^K x_\tau^k$ and $\overline{Y}_\tau = \sum_{k=1}^K y_\tau^k$, respectively.

Now, building on the seminal work on aggregation with optimization by Koopmans (1957), consider a *group input requirement correspondence* for period τ defined as the Minkowski sum[8] of the individual input requirement correspondences for a given period τ[9]:

$$\overline{L}_\tau(Y) \equiv \sum_{\oplus k=1}^{K} L_\tau^k\left(y^k\right),\ y^k \in \mathbb{R}_+^M,\ k = 1, \ldots, K. \tag{19}$$

[8]The symbol $\oplus$ is used following a common notation in mathematics to distinguish the summation of sets (or Minkowski summation) from the standard summation. See Oks and Sharir (2006) for more details on Minkowski summation.

[9]This follows Färe et al. (2004), which in turn has built upon the work of Koopmans (1957) and Färe and Zelenyuk (2003), with extensions here to the intertemporal context.

This shows the possible overall group input requirement sets that would allow a given output level Y by the organizations (i.e., output production cannot be reallocated among organizations, a restriction we will relax later). Based on this definition, the *group cost function* can be defined analogously to the individual cost function, as:

$$\begin{aligned} \overline{C}_\tau(Y,w) \equiv \inf_x \left\{ wx : x \in \overline{L}_\tau(Y) \right\}, \\ y^k \in \mathbb{R}^M_+,\ k = 1,\ldots,K,\ w \in \mathbb{R}^N_{++}, \end{aligned} \tag{20}$$

with the *group cost efficiency* measure defined as

$$\overline{\mathrm{CE}}_\tau\left(Y,\overline{X},w\right) \equiv \frac{\overline{C}_\tau(Y,w)}{w\overline{X}}, \quad \text{for } w\overline{X} \neq 0. \tag{21}$$

Note the assumption here that the input information on organization is aggregated using the same input prices. We can think of this common price as a theoretical benchmark (e.g., equilibrium) price against which group cost efficiency is derived. Conceptually, this is similar to the assumption that the group technical efficiency is benchmarked against common technology although in practice organizations can use different technologies. For practical applications, cost/revenue and quantity data is commonly used to construct average prices which can be treated as the common price (see, e.g., Fukuyama and Weber 2008). Another option is to derive shadow prices after requiring the 'Law of One Price' (see, e.g., Kuosmanen et al. 2006, 2010). The limitations of different methods of estimating the prices are something researchers need to consider for their own applications, as they already do for alternative estimators of the technology.

In order to determine the aggregation functions, below we summarize a number of key aggregation results.

Lemma 1 *We have*

$$\overline{C}_\tau\left(Y_\tau, w_j\right) = \sum_{k=1}^{K} C^k_\tau\left(y^k_\tau, w_j\right),\ y^k_\tau \in \mathbb{R}^M_+,\ w_j \in \mathbb{R}^N_{++}, \tag{22}$$

and so, the group cost efficiency *is:*

$$\overline{\mathrm{CE}}_\tau \equiv \overline{\mathrm{CE}}_\tau\left(Y_\tau, \overline{X}_j, w_j\right) = \sum_{k=1}^{K} \mathrm{CE}_\tau\left(y^k_\tau, x^k_j, w_j\right) \cdot W^k_j, \tag{23}$$

with weights

$$W_j^k = \frac{w_j x_j^k}{w_j \overline{X}_j}, \quad k = 1, \ldots, K, \tag{24}$$

so that $\overline{\mathrm{CE}}_\tau$ *can be further decomposed as*

$$\overline{\mathrm{CE}}_\tau\left(Y_\tau, \overline{X}_j, w_j\right) = \overline{\mathrm{ITE}}_\tau\left(Y_\tau, \overline{X}_j\right) \times \overline{\mathrm{IAE}}_\tau\left(Y_\tau, \overline{X}_j, w_j\right), \text{ for all } \tau, j, \tag{25}$$

where

$$\overline{\mathrm{ITE}}_\tau \equiv \overline{\mathrm{ITE}}_\tau\left(Y_\tau, \overline{X}_j\right) = \sum_{k=1}^{K} \mathrm{ITE}_\tau^k\left(y_\tau^k, x_j^k\right) \cdot W_j^k, \quad k = 1, \ldots, K, \tag{26}$$

represents input-oriented group technical efficiency, *while*

$$\overline{\mathrm{IAE}}_\tau \equiv \overline{\mathrm{IAE}}_\tau\left(Y_\tau, \overline{X}_j, w_j\right) = \sum_{k=1}^{K} \mathrm{IAE}_\tau^k\left(y_\tau^k, x_j^k, w_j\right) \cdot W_{ae,\tau,j}^k, \tag{27}$$

represents input-oriented group allocative efficiency, *with the weights given by*[10]

$$W_{ae,\tau,j}^k = \frac{w_j x_j^k \mathrm{ITE}_\tau^k\left(y_\tau^k, x_j^k\right)}{\sum_{k=1}^{K} w_j x_j^k \mathrm{ITE}_\tau^k\left(y_\tau^k, x_j^k\right)}, \quad k = 1, \ldots, K. \tag{28}$$

In words, this lemma shows that the minimum overall cost of the organizations considered as a group equals the sum of their individual minimum costs, given their individual output production and facing the common input prices. We can then use a weighted sum of individual efficiencies to express group efficiencies, while also preserving the decomposition into technical and allocative efficiency on the aggregate level (25), analogous to the one observed

[10]Note that the weights here involve the technical efficiency scores. This is intuitive because while technical efficiency is measured from some base point x, allocative efficiency is not measured from this point x but from its radial projection on to the frontier, defined by adjusting (multiplying) x by the relevant technical efficiency score to bring it to the frontier. Thus, it is intuitive that this projection (i.e., $\widehat{x} = x\mathrm{ITE}\,(y, x)$ rather than x) must be involved in the aggregation of the allocative efficiency (evaluated at some relevant price level w). This explains why technical efficiency appears in the weights of aggregation of allocative efficiency.

on the individual level. Note that these weights are not ad hoc but are derived from cost-minimizing behavior w.r.t. the aggregation structure (19) and relative to common input prices. Also note that throughout, τ and j are two time periods (which can be the same). The proof of Lemma 1 is similar to that of the output-oriented case in Färe and Zelenyuk (2003) and is therefore omitted (see Färe et al. (2004) for some details).[11]

This lemma takes the allocation among organizations of required outputs as given and considers the overall group level of inputs; but where technology does not exhibit constant returns to scale (or some organizations have inferior technologies), further output may be possible through reallocating output production between organizations. These gains would be additional to those from all organizations operating efficiently with respect to their current production requirements.

3.2 Aggregate Efficiency Measures Allowing Reallocation

In order to measure the gains from relaxing the restriction on reallocation of resources between organizations in a group, consider a *group potential technology*, as a Minkowski sum of individual technologies for a given period τ[12]:

$$T_\tau^* \equiv \sum_{\oplus k=1}^{K} T_\tau^k. \tag{29}$$

By aggregating technology sets instead of output and input requirement correspondences, this group potential technology allows full reallocation of inputs and outputs among organizations in the group. An equivalent characterization is the *group potential input requirement correspondence*, defined as

$$L_\tau^*\left(\overline{Y}\right) = \left\{x : (x, \overline{Y}) \in T_\tau^*\right\}. \tag{30}$$

Now, using this group technology, let the *group potential input-oriented technical efficiency* be defined as

[11]This result can also be viewed as a cost-analogue of Koopmans' theorem of aggregation of profit functions (Koopmans 1957). For a related result in the context of the consumer analysis area of economic theory, see Luenberger (1996).

[12]This technology aggregation structure was earlier used in Li and Ng (1995), Blackorby and Russell (1999), and Nesterenko and Zelenyuk (2007).

$$\mathrm{ITE}^*_\tau = \mathrm{ITE}^*_\tau\left(\overline{Y}_\tau, \overline{X}_j\right) = \inf_\lambda \left\{\lambda : \lambda \overline{X}_j \in L^*_\tau\left(\overline{Y}_\tau\right)\right\}. \tag{31}$$

while the dual characterization of $L^*_\tau(\overline{Y})$, the *group potential cost function* can be defined as

$$C^*_\tau\left(\overline{Y}_\tau, w_j\right) = \inf_x \left\{w_j x : x \in L^*_\tau\left(\overline{Y}_\tau\right)\right\}. \tag{32}$$

and the related *group potential cost efficiency* is then defined as

$$\mathrm{CE}^*_\tau = \mathrm{CE}^*_\tau\left(\overline{Y}_\tau, \overline{X}_j, w_j\right) = \frac{C^*_\tau\left(\overline{Y}_\tau, w_j\right)}{w_j \overline{X}_j},\ w_j \overline{X}_j \neq 0. \tag{33}$$

Note that $\mathrm{CE}^*_\tau \leq \mathrm{ITE}^*_\tau$, and so analogous to the individual level, we define the *group potential input-oriented allocative efficiency*, which closes the above inequality:

$$\mathrm{CE}^*_\tau\left(\overline{Y}_\tau, \overline{X}_j, w_j\right) = \mathrm{ITE}^*_\tau\left(\overline{Y}_\tau, \overline{X}_j\right) \times \mathrm{IAE}^*_\tau\left(\overline{Y}_\tau, \overline{X}_j, w_j\right), \text{ for all } \tau, j. \tag{34}$$

In words, (31) and (33) measure group efficiency relative to the group potential input requirement correspondence (30), in a manner analogous to measurements on the individual level.

In comparing the group input requirement correspondence when full reallocation is permitted to that when it is not, the following important lemma emerges:

Lemma 2 *We have*

$$\overline{L}_\tau(Y_\tau) \subseteq L^*_\tau\left(\overline{Y}_\tau\right). \tag{35}$$

A proof of (35) is analogous to that for the output-oriented case found in Nesterenko and Zelenyuk (2007) and so we omit it. Intuitively, the group potential input requirement correspondence (where full reallocation is permitted) will always encompass a linear aggregation of the input requirement correspondences (where full reallocation is not permitted). By implication, for any $\left(Y_\tau, \overline{Y}_\tau, w_j\right)$ we have $C^*_\tau\left(\overline{Y}_\tau, w_j\right) \leq \overline{C}_\tau\left(Y_\tau, w_j\right)$ from which it follows that $\mathrm{CE}^*_\tau \leq \overline{\mathrm{CE}}_\tau$. Analogous to the terminology of Nesterenko and Zelenyuk (2007), we can define the *group cost reallocative efficiency*, as the multiplicative residual which closes the latter inequality, i.e.,

$$\mathrm{CRE}^*_\tau = \mathrm{CRE}^*_\tau\left(Y_\tau, \overline{Y}_\tau, w_j\right) = C^*_\tau\left(\overline{Y}_\tau, w_j\right)/\overline{C}_\tau\left(Y_\tau, w_j\right), \tag{36}$$

and so we get the following decomposition of the group cost efficiency:

$$\mathrm{CE}^*_\tau = \overline{\mathrm{CE}}_\tau \times \mathrm{CRE}^*_\tau, \text{ for all } \tau. \tag{37}$$

Moreover, CRE^*_τ can be further decomposed, as summarized in the next lemma.

Lemma 3 *We have*

$$\mathrm{CRE}^*_\tau = \mathrm{ITRE}^*_\tau \times \mathrm{IARE}^*_\tau, \text{ for all } \tau, \tag{38}$$

where group input-oriented technical reallocative efficiency *is:*

$$\mathrm{ITRE}^*_\tau = \mathrm{ITE}^*_\tau/\overline{\mathrm{ITE}}_\tau, \tag{39}$$

and group input-oriented allocative reallocative efficiency *is:*

$$\mathrm{IARE}^*_\tau = \mathrm{IAE}^*_\tau/\overline{\mathrm{IAE}}_\tau. \tag{40}$$

In words, this lemma says that each reallocative efficiency measure reveals the difference for the group between individual efficiency in each organization in input orientation (w.r.t. their individual output plans), and the collective efficiency in input orientation, where outputs are permitted to be reallocated among organizations.

With the same logic as Nesterenko and Zelenyuk (2007), we can define reallocative measures for individual input-oriented organizations as:

$$\mathrm{CRE}^k_\tau = \mathrm{CE}^*_\tau/\mathrm{CE}^k_\tau, \tag{41}$$

$$\mathrm{ITRE}^k_\tau = \mathrm{ITE}^*_\tau/\mathrm{ITE}^k_\tau, \tag{42}$$

$$\mathrm{IARE}^k_\tau = \mathrm{IAE}^*_\tau/\mathrm{IAE}^k_\tau. \tag{43}$$

Next, in Lemma 4, we summarize the relationship between individual and group reallocative measures.

Lemma 4 *We have*

$$\mathrm{CRE}_{\tau}^{*}=\left(\sum_{k=1}^{K}\left(\mathrm{CRE}_{\tau}^{k}\left(y_{\tau}^{k},x_{j}^{k},w_{j}\right)\right)^{-1}\cdot W_{j}^{k}\right)^{-1}, \tag{44}$$

$$\mathrm{ITRE}_{\tau}^{*}=\left(\sum_{k=1}^{K}\left(\mathrm{ITRE}_{\tau}^{k}\left(y_{\tau}^{k},x_{j}^{k}\right)\right)^{-1}\cdot W_{j}^{k}\right)^{-1}, \tag{45}$$

$$\mathrm{IARE}_{\tau}^{*}=\left(\sum_{k=1}^{K}\left(\mathrm{IARE}_{\tau}^{k}\left(y_{\tau}^{k},x_{j}^{k},w_{j}\right)\right)^{-1}\cdot W_{ae,\tau,j}^{k}\right)^{-1}, \tag{46}$$

where (24) *and* (28) *define the weights.*

In turn, note that decompositions (25), (37), and (38) imply the following decomposition of group potential cost efficiency:

$$\mathrm{CE}_{\tau}^{*}=\overline{\mathrm{ITE}}_{\tau}\times\overline{\mathrm{IAE}}_{\tau}\times\mathrm{ITRE}_{\tau}^{*}\times\mathrm{IARE}_{\tau}^{*},\ \text{for all } \tau. \tag{47}$$

Similar results hold for output orientation, with group efficiency measures $\left(\overline{\mathrm{RE}}_{\tau},\overline{\mathrm{OTE}}_{\tau},\overline{\mathrm{OAE}}_{\tau}\right)$, group potential efficiency measures $\left(\mathrm{RE}_{\tau}^{*},\mathrm{OTE}_{\tau}^{*},\mathrm{OAE}_{\tau}^{*}\right)$, and group reallocative efficiency measures $\left(\mathrm{RRE}_{\tau}^{*},\mathrm{OTRE}_{\tau}^{*},\mathrm{OARE}_{\tau}^{*}\right)$ for revenue, output-oriented technical and output-oriented allocative efficiencies, respectively, defined analogously to the input-oriented results stated above (see Mayer and Zelenyuk 2014a). Having restated the key results for aggregate efficiency measures, with and without reallocation, we can now construct our aggregate HMPIs, as we do in the following section. We aim to do so in a manner which permits decompositions similar to (47).

4 Aggregate Hicks-Moorsteen Productivity Indices

Here, we start by constructing a group potential profitability HMPI in terms of the group potential revenue and cost efficiencies (i.e., an aggregate profitability HMPI allowing full reallocation), defined similarly to the individual profitability HMPI, and then decompose it into technical and allocative components.

Proposition 1 *Let the group potential profitability Hicks-Moorsteen productivity index from period* s *to* t *be given by*

$$\begin{aligned}\mathrm{PHM}^{*}_{st} &= \mathrm{PHM}^{*}_{st}\left(\overline{Y}_s,\overline{Y}_t,\overline{X}_s,\overline{X}_t,p_s,p_t,w_s,w_t\right)\\ &= \left[\left(\frac{\frac{\mathrm{RE}^{*}_{s}\left(\overline{X}_s,\overline{Y}_t,p_t\right)}{\mathrm{RE}^{*}_{s}\left(\overline{X}_s,\overline{Y}_s,p_s\right)}}{\frac{\mathrm{CE}^{*}_{s}\left(\overline{Y}_s,\overline{X}_t,w_t\right)}{\mathrm{CE}^{*}_{s}\left(\overline{Y}_s,\overline{X}_s,w_s\right)}}\times\frac{\frac{\mathrm{RE}^{*}_{t}\left(\overline{X}_t,\overline{Y}_t,p_t\right)}{\mathrm{RE}^{*}_{t}\left(\overline{X}_t,\overline{Y}_s,p_s\right)}}{\frac{\mathrm{CE}^{*}_{t}\left(\overline{Y}_t,\overline{X}_t,w_t\right)}{\mathrm{CE}^{*}_{t}\left(\overline{Y}_t,\overline{X}_s,w_s\right)}}\right)^{-1}\right]^{1/2},\end{aligned} \tag{48}$$

then it can be decomposed to technical and allocative components for any s, t *as*

$$\mathrm{PHM}^{*}_{st} = \mathrm{HM}^{*}_{st}\times\mathrm{AHM}^{*}_{st}, \tag{49}$$

where

$$\begin{aligned}\mathrm{HM}^{*}_{st} &= \mathrm{HM}^{*}_{st}\left(\overline{Y}_s,\overline{Y}_t,\overline{X}_s,\overline{X}_t\right)\\ &= \left[\left(\frac{\frac{\mathrm{OTE}^{*}_{s}\left(\overline{X}_s,\overline{Y}_t\right)}{\mathrm{OTE}^{*}_{s}\left(\overline{X}_s,\overline{Y}_s\right)}}{\frac{\mathrm{ITE}^{*}_{s}\left(\overline{Y}_s,\overline{X}_t\right)}{\mathrm{ITE}^{*}_{s}\left(\overline{Y}_s,\overline{X}_s\right)}}\times\frac{\frac{\mathrm{OTE}^{*}_{t}\left(\overline{X}_t,\overline{Y}_t\right)}{\mathrm{OTE}^{*}_{t}\left(\overline{X}_t,\overline{Y}_s\right)}}{\frac{\mathrm{ITE}^{*}_{t}\left(\overline{Y}_t,\overline{X}_t\right)}{\mathrm{ITE}^{*}_{t}\left(\overline{Y}_t,\overline{X}_s\right)}}\right)^{-1}\right]^{1/2},\end{aligned} \tag{50}$$

is the group potential Hicks-Moorsteen productivity index from s *to* t*, and*

$$\begin{aligned}\mathrm{AHM}^{*}_{st} &= \mathrm{AHM}^{*}_{st}\left(\overline{Y}_s,\overline{Y}_t,\overline{X}_s,\overline{X}_t,p_s,p_t,w_s,w_t\right)\\ &= \left[\left(\frac{\frac{\mathrm{OAE}^{*}_{s}\left(\overline{X}_s,\overline{Y}_t,p_t\right)}{\mathrm{OAE}^{*}_{s}\left(\overline{X}_s,\overline{Y}_s,p_s\right)}}{\frac{\mathrm{IAE}^{*}_{s}\left(\overline{Y}_s,\overline{X}_t,w_t\right)}{\mathrm{IAE}^{*}_{s}\left(\overline{Y}_s,\overline{X}_s,w_s\right)}}\times\frac{\frac{\mathrm{OAE}^{*}_{t}\left(\overline{X}_t,\overline{Y}_t,p_t\right)}{\mathrm{OAE}^{*}_{t}\left(\overline{X}_t,\overline{Y}_s,p_s\right)}}{\frac{\mathrm{IAE}^{*}_{t}\left(\overline{Y}_t,\overline{X}_t,w_t\right)}{\mathrm{IAE}^{*}_{t}\left(\overline{Y}_t,\overline{X}_s,w_s\right)}}\right)^{-1}\right]^{1/2},\end{aligned} \tag{51}$$

is the group potential allocative Hicks-Moorsteen productivity index from s *to* t.

To demonstrate this proposition, substitute decompositions of the group potential revenue and cost efficiency measures, (34) and its output-oriented analogue into the group potential profitability HMPI (for each period), and afterward separate out the group potential primal and allocative HMPI measures. Note that these measures are in the same form as the individual HMPIs.

Intuitively, these group potential HMPIs capture the productivity change for the group between the two periods, allowing full reallocation of outputs and inputs among the organizations. Improvements in these measures indicate that the group potential productivity (i.e., productivity when full reallocation

is possible) has improved. This measure is particularly relevant in those cases where such reallocation is possible—for studying a firm with many branches, countries forming an economic union where such reallocation is relevant, etc.

Our goal now is to relate aggregate HMPI measures to the individual HMPIs. To achieve this, we decompose these group potential HMPIs into the productivity change with and without allowing full reallocation, the latter of which can be related to the individual measures. We present these in the next two propositions and then show their relationship to the group potential HMPIs as corollaries.

Proposition 2 *Let the group profitability Hicks-Moorsteen productivity index from* s *to* t *be given by*

$$\begin{aligned}\overline{\mathrm{PHM}}_{st} &= \overline{\mathrm{PHM}}_{st}(Y_s, Y_t, X_s, X_t, p_s, p_t, w_s, w_t) \\ &= \left[\frac{\sum_{k=1}^{K} \mathrm{RE}_s\left(x_s^k, y_t^k, p_t\right)\cdot S_t^k}{\sum_{k=1}^{K} \mathrm{RE}_s\left(x_s^k, y_s^k, p_s\right)\cdot S_s^k} \Big/ \frac{\sum_{k=1}^{K} \mathrm{CE}_s\left(y_s^k, x_t^k, w_t\right)\cdot W_t^k}{\sum_{k=1}^{K} \mathrm{CE}_s\left(y_s^k, x_s^k, w_s\right)\cdot W_s^k}\right]^{-1/2} \\ &\quad \times \left[\frac{\sum_{k=1}^{K} \mathrm{RE}_t\left(x_t^k, y_t^k, p_t\right)\cdot S_t^k}{\sum_{k=1}^{K} \mathrm{RE}_t\left(x_t^k, y_s^k, p_s\right)\cdot S_s^k} \Big/ \frac{\sum_{k=1}^{K} \mathrm{CE}_t\left(y_t^k, x_t^k, w_t\right)\cdot W_t^k}{\sum_{k=1}^{K} \mathrm{CE}_t\left(y_t^k, x_s^k, w_s\right)\cdot W_s^k}\right]^{-1/2}\end{aligned} \tag{52}$$

then it can be decomposed into technical and allocative components (for any s, t*) as*

$$\overline{\mathrm{PHM}}_{st} = \overline{\mathrm{HM}}_{st} \times \overline{\mathrm{AHM}}_{st}, \tag{53}$$

where

$$\begin{aligned}\overline{\mathrm{HM}}_{st} &= \overline{\mathrm{HM}}_{st}(Y_s, Y_t, X_s, X_t) \\ &= \left[\frac{\sum_{k=1}^{K} \mathrm{OTE}_s\left(x_s^k, y_t^k\right)\cdot S_t^k}{\sum_{k=1}^{K} \mathrm{OTE}_s\left(x_s^k, y_s^k\right)\cdot S_s^k} \Big/ \frac{\sum_{k=1}^{K} \mathrm{ITE}_s\left(y_s^k, x_t^k\right)\cdot W_t^k}{\sum_{k=1}^{K} \mathrm{ITE}_s\left(y_s^k, x_s^k\right)\cdot W_s^k}\right]^{-1/2} \\ &\quad \times \left[\frac{\sum_{k=1}^{K} \mathrm{OTE}_t\left(x_t^k, y_t^k\right)\cdot S_t^k}{\sum_{k=1}^{K} \mathrm{OTE}_t\left(x_t^k, y_s^k\right)\cdot S_s^k} \Big/ \frac{\sum_{k=1}^{K} \mathrm{ITE}_t\left(y_t^k, x_t^k\right)\cdot W_t^k}{\sum_{k=1}^{K} \mathrm{ITE}_t\left(y_t^k, x_s^k\right)\cdot W_s^k}\right]^{-1/2},\end{aligned} \tag{54}$$

is the group Hicks-Moorsteen productivity index from s *to* t*, and*

$$\overline{\mathrm{AHM}}_{st} = \overline{\mathrm{AHM}}_{st}(Y_s, Y_t, X_s, X_t, p_s, p_t, w_s, w_t)$$
$$= \left[\frac{\sum_{k=1}^{K} \mathrm{OAE}_s\left(x_s^k, y_t^k, p_t\right)\cdot S_{ae,s,t}^k}{\sum_{k=1}^{K} \mathrm{OAE}_s\left(x_s^k, y_s^k, p_s\right)\cdot S_{ae,s,s}^k}\Bigg/\frac{\sum_{k=1}^{K} \mathrm{IAE}_s\left(y_s^k, x_t^k, w_t\right)\cdot W_{ae,s,t}^k}{\sum_{k=1}^{K} \mathrm{IAE}_s\left(y_s^k, x_s^k, w_s\right)\cdot W_{ae,s,s}^k}\right]^{-1/2} \times \left[\frac{\sum_{k=1}^{K} \mathrm{OAE}_t\left(x_t^k, y_t^k, p_t\right)\cdot S_{ae,t,t}^k}{\sum_{k=1}^{K} \mathrm{OAE}_t\left(x_t^k, y_s^k, p_s\right)\cdot S_{ae,t,s}^k}\Bigg/\frac{\sum_{k=1}^{K} \mathrm{IAE}_t\left(y_t^k, x_t^k, w_t\right)\cdot W_{ae,t,t}^k}{\sum_{k=1}^{K} \mathrm{IAE}_t\left(y_t^k, x_s^k, w_s\right)\cdot W_{ae,t,s}^k}\right]^{-1/2} \tag{55}$$

is the group allocative Hicks-Moorsteen productivity index from s *to* t, where the weights for the output orientation part are the revenue analogues of the cost-shares derived in the previous section (see Mayer and Zelenyuk (2014a, b) for the details).

Again, the proof of this proposition follows from taking the group profitability HMPI, substituting in the decompositions of the group revenue and cost efficiency measures, (25) and its output-oriented analogue, for each period, and rearranging so as to isolate the group primal and allocative HMPI measures. Note that for the group profitability and primal HMPIs, the weights for the measure of each orientation (input or output) are calculated using that factor-price combination for that period (e.g., input-oriented measures use the weights for the same period as the inputs and prices). For the group allocative HMPI, the weights depend both on the period of the technology and the period of the variable factors (cost shares for input orientation, revenue shares for output orientation).

Intuitively, each of the group HMPIs measures the productivity change of the overall group taking current input endowments (for output-oriented measures) and output production (for input-oriented measures) as given, that is, without allowing full reallocation. Increases in these measures indicate that the group productivity (taking the input or output allocation among organizations as given) has improved. They have each been constructed in terms of the group efficiency measures, which in turn are constructed from the individual efficiency measures (with appropriate weights). It is the latter that are usually estimated in practice, and this result shows how these individual measures can be consistently aggregated into a group productivity index. Moreover, the aggregation results of Färe and Zelenyuk (2003) for the group efficiency measures decompose analogously to the individual measures (following (25)), and so our group productivity indices also

decompose analogously to the individual productivity indices, (17). Thus, the group profitability HMPI can be decomposed following (53), i.e., into the group productivity change due to changes in group efficiency or technology (the group primal HMPI) and group productivity change due to changes in the allocation of factors within each organization in the group or changes in the prices faced by the group (the group allocative HMPI). One can also obtain similar results for the group reallocative HMPIs, as summarized next.

Proposition 3 *Let the group profitability reallocative Hicks-Moorsteen productivity index from* s *to* t *be given by*

$$\begin{aligned}
\mathrm{PRHM}^*_{st} &= \mathrm{PRHM}^*_{st}\left(\overline{Y}_s, \overline{Y}_t, \overline{X}_s, \overline{X}_t, Y_s, Y_t, X_s, X_t, p_s, p_t, w_s, w_t\right) \\
&= \left[\left(\frac{\mathrm{RRE}^*_s\left(\overline{X}_s, X_s, \overline{Y}_t, Y_t, p_t\right)/\mathrm{RRE}^*_s\left(\overline{X}_s, X_s, \overline{Y}_s, Y_s, p_s\right)}{\mathrm{CRE}^*_s\left(\overline{Y}_s, Y_s, \overline{X}_t, X_t, w_t\right)/\mathrm{CRE}^*_s\left(\overline{Y}_s, Y_s, \overline{X}_s, X_s, w_s\right)}\right)^{-1}\right]^{1/2} \\
&\quad \times \left[\left(\frac{\mathrm{RRE}^*_t\left(\overline{X}_t, X_t, \overline{Y}_t, Y_t, p_t\right)/\mathrm{RRE}^*_t\left(\overline{X}_t, X_t, \overline{Y}_s, Y_s, p_s\right)}{\mathrm{CRE}^*_t\left(\overline{Y}_t, Y_t, \overline{X}_t, X_t, w_t\right)/\mathrm{CRE}^*_t\left(\overline{Y}_t, Y_t, \overline{X}_s, X_s, w_s\right)}\right)^{-1}\right]^{1/2} \\
&= \left[\frac{\left(\sum_{k=1}^{K}\left(\mathrm{RRE}^k_s\left(x^k_s, y^k_t, p_t\right)\right)^{-1}\cdot S^k_t\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{RRE}^k_s\left(x^k_s, y^k_s, p_s\right)\right)^{-1}\cdot S^k_s\right)^{-1}}\Bigg/\frac{\left(\sum_{k=1}^{K}\left(\mathrm{CRE}^k_s\left(y^k_s, x^k_t, w_t\right)\right)^{-1}\cdot W^k_t\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{CRE}^k_s\left(y^k_s, x^k_s, w_s\right)\right)^{-1}\cdot W^k_s\right)^{-1}}\right]^{1/2} \\
&\quad \times \left[\frac{\left(\sum_{k=1}^{K}\left(\mathrm{RRE}^k_t\left(x^k_t, y^k_t, p_t\right)\right)^{-1}\cdot S^k_t\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{RRE}^k_t\left(x^k_t, y^k_s, p_s\right)\right)^{-1}\cdot S^k_s\right)^{-1}}\Bigg/\frac{\left(\sum_{k=1}^{K}\left(\mathrm{CRE}^k_t\left(y^k_t, x^k_t, w_t\right)\right)^{-1}\cdot W^k_t\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{CRE}^k_t\left(y^k_t, x^k_s, w_s\right)\right)^{-1}\cdot W^k_s\right)^{-1}}\right]^{1/2},
\end{aligned} \tag{56}$$

then for any s *and* t*, it can be decomposed as*

$$\mathrm{PRHM}^*_{st} = \mathrm{RHM}^*_{st} \times \mathrm{ARHM}^*_{st}, \tag{57}$$

where

$$\begin{aligned}
\mathrm{RHM}^{*}_{st} &= \mathrm{RHM}^{*}_{st}(\overline{Y}_s, \overline{Y}_t, \overline{X}_s, \overline{X}_t, Y_s, Y_t, X_s, X_t) \\
&= \left[\left(\frac{\mathrm{OTRE}^{*}_{s}\left(\overline{X}_s, X_s, \overline{Y}_t, Y_t\right)/\mathrm{OTRE}^{*}_{s}\left(\overline{X}_s, X_s, \overline{Y}_s, Y_s\right)}{\mathrm{ITRE}^{*}_{s}\left(\overline{Y}_s, Y_s, \overline{X}_t, X_t\right)/\mathrm{ITRE}^{*}_{s}\left(\overline{Y}_s, Y_s, \overline{X}_s, X_s\right)}\right)^{-1}\right]^{1/2} \\
&\quad \times \left[\left(\frac{\mathrm{OTRE}^{*}_{t}\left(\overline{X}_t, X_t, \overline{Y}_t, Y_t\right)/\mathrm{OTRE}^{*}_{t}\left(\overline{X}_t, X_t, \overline{Y}_s, Y_s\right)}{\mathrm{ITRE}^{*}_{t}\left(\overline{Y}_t, Y_t, \overline{X}_t, X_t\right)/\mathrm{ITRE}^{*}_{t}\left(\overline{Y}_t, Y_t, \overline{X}_s, X_s\right)}\right)^{-1}\right]^{1/2} \\
&= \left[\frac{\left(\sum_{k=1}^{K}\left(\mathrm{OTRE}^{k}_{s}\left(x^{k}_{s}, y^{k}_{t}\right)\right)^{-1}\cdot S^{k}_{t}\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{OTRE}^{k}_{s}\left(x^{k}_{s}, y^{k}_{s}\right)\right)^{-1}\cdot S^{k}_{s}\right)^{-1}}\Bigg/\frac{\left(\sum_{k=1}^{K}\left(\mathrm{ITRE}^{k}_{s}\left(y^{k}_{s}, x^{k}_{t}\right)\right)^{-1}\cdot W^{k}_{t}\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{ITRE}^{k}_{s}\left(y^{k}_{s}, x^{k}_{s}\right)\right)^{-1}\cdot W^{k}_{s}\right)^{-1}}\right]^{1/2} \\
&\quad \times \left[\frac{\left(\sum_{k=1}^{K}\left(\mathrm{OTRE}^{k}_{t}\left(x^{k}_{t}, y^{k}_{t}\right)\right)^{-1}\cdot S^{k}_{t}\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{OTRE}^{k}_{t}\left(x^{k}_{t}, y^{k}_{s}\right)\right)^{-1}\cdot S^{k}_{s}\right)^{-1}}\Bigg/\frac{\left(\sum_{k=1}^{K}\left(\mathrm{ITRE}^{k}_{t}\left(y^{k}_{t}, x^{k}_{t}\right)\right)^{-1}\cdot W^{k}_{t}\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{ITRE}^{k}_{t}\left(y^{k}_{t}, x^{k}_{s}\right)\right)^{-1}\cdot W^{k}_{s}\right)^{-1}}\right]^{1/2},
\end{aligned} \tag{58}$$

is the group reallocative Hicks-Moorsteen productivity index from s *to* t*, and*

$$\begin{aligned}
\mathrm{ARHM}^{*}_{st} &= \mathrm{ARHM}^{*}_{st}\left(\overline{Y}_s, \overline{Y}_t, \overline{X}_s, \overline{X}_t, Y_s, Y_t, X_s, X_t, p_s, p_t, w_s, w_t\right) \\
&= \left[\left(\frac{\mathrm{OARE}^{*}_{s}\left(\overline{X}_s, X_s, \overline{Y}_t, Y_t, p_t\right)/\mathrm{OARE}^{*}_{s}\left(\overline{X}_s, X_s, \overline{Y}_s, Y_s, p_s\right)}{\mathrm{IARE}^{*}_{s}\left(\overline{Y}_s, Y_s, \overline{X}_t, X_t, w_t\right)/\mathrm{IARE}^{*}_{s}\left(\overline{Y}_s, Y_s, \overline{X}_s, X_s, w_s\right)}\right)^{-1}\right]^{1/2} \\
&\quad \times \left[\left(\frac{\mathrm{OARE}^{*}_{t}\left(\overline{X}_t, X_t, \overline{Y}_t, Y_t, p_t\right)/\mathrm{OARE}^{*}_{t}\left(\overline{X}_t, X_t, \overline{Y}_s, Y_s, p_s\right)}{\mathrm{IARE}^{*}_{t}\left(\overline{Y}_t, Y_t, \overline{X}_t, X_t, w_t\right)/\mathrm{IARE}^{*}_{t}\left(\overline{Y}_t, Y_t, \overline{X}_s, X_s, w_s\right)}\right)^{-1}\right]^{1/2} \\
&= \left[\frac{\left(\sum_{k=1}^{K}\left(\mathrm{OARE}^{k}_{s}\left(x^{k}_{s}, y^{k}_{t}, p_t\right)\right)^{-1}\cdot S^{k}_{ae,s,t}\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{OARE}^{k}_{s}\left(x^{k}_{s}, y^{k}_{s}, p_s\right)\right)^{-1}\cdot S^{k}_{ae,s,s}\right)^{-1}}\Bigg/\frac{\left(\sum_{k=1}^{K}\left(\mathrm{IARE}^{k}_{s}\left(y^{k}_{s}, x^{k}_{t}, w_t\right)\right)^{-1}\cdot W^{k}_{ae,s,t}\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{IARE}^{k}_{s}\left(y^{k}_{s}, x^{k}_{s}, w_s\right)\right)^{-1}\cdot W^{k}_{ae,s,s}\right)^{-1}}\right]^{1/2} \\
&\quad \times \left[\frac{\left(\sum_{k=1}^{K}\left(\mathrm{OARE}^{k}_{t}\left(x^{k}_{t}, y^{k}_{t}, p_t\right)\right)^{-1}\cdot S^{k}_{ae,t,t}\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{OARE}^{k}_{t}\left(x^{k}_{t}, y^{k}_{s}, p_s\right)\right)^{-1}\cdot S^{k}_{ae,t,s}\right)^{-1}}\Bigg/\frac{\left(\sum_{k=1}^{K}\left(\mathrm{IARE}^{k}_{t}\left(y^{k}_{t}, x^{k}_{t}, w_t\right)\right)^{-1}\cdot W^{k}_{ae,t,t}\right)^{-1}}{\left(\sum_{k=1}^{K}\left(\mathrm{IARE}^{k}_{t}\left(y^{k}_{t}, x^{k}_{s}, w_s\right)\right)^{-1}\cdot W^{k}_{ae,t,s}\right)^{-1}}\right]^{1/2},
\end{aligned} \tag{59}$$

is the group allocative reallocative Hicks-Moorsteen productivity index from s *to* t.

Intuitively, this proposition says that the change in productivity from permitting full reallocation of inputs and outputs within the group of organizations is measured by the group reallocative HMPIs, a change in addition to that of every organization operating efficiently. Similarly to as we did before, this proposition can be proved by starting with the group profitability reallocative HMPI, substituting in the decompositions of the group revenue and cost reallocative efficiency measures, (38) and its output-oriented analogue, for each period, and then rearranging to get the desired components. For each measure, the last equality (expressing it in terms of individual reallocative efficiency measures) follows from Lemma 4.

Furthermore, the group potential profitability HMPI can decompose into the group profitability HMPI and the group profitability reallocative HMPI, as done in the next corollary.

Corollary 1 *We have*

$$\mathrm{PHM}^{*}_{st} = \overline{\mathrm{PHM}}_{st} \times \mathrm{PRHM}^{*}_{st}, \tag{60}$$

for any two periods s *and* t.

This proposition is derived by starting with the group potential profitability HMPI (48), substituting in the decompositions of group potential revenue and cost efficiency, (37) and its output-oriented analogue, and rearranging using Propositions (2) and (3).

This decomposition (60) is important because it shows the source of changes in group potential productivity changes. Consider an improvement in group potential productivity—group productivity could improve proportionally, in the case of technological improvement (here, group reallocative productivity would be close to unity); or group productivity may be unchanged, if the new allocation merely shifts along the frontier (here, group reallocative productivity would proportionally increase); or group productivity could end up lower, where individual efficiency is sacrificed to move toward an optimal group allocation (here, group potential productivity would not increase as much as group reallocative productivity); or it could be the case that both group productivity and group reallocative productivity improve together.

Similar decompositions hold for the group potential primal and allocative HMPIs, as we summarise in the next corollary.

Corollary 2 *We have*

$$\mathrm{HM}^{*}_{st} = \overline{\mathrm{HM}}_{st} \times \mathrm{RHM}^{*}_{st}, \tag{61}$$

and

$$\mathrm{AHM}^{*}_{st} = \overline{\mathrm{AHM}}_{st} \times \mathrm{ARHM}^{*}_{st}, \tag{62}$$

both for any periods s *and* t.

To prove this result, given the group potential [allocative] HMPI from (50), substitute in the decompositions of group potential technical [allocative] efficiency, (39) [(40)] and their output-oriented analogues, before rearranging using Propositions (2) and (3).

The decomposition of the primal group potential HMPI, (61), is of particular value to researchers because it does not require the price information that is necessary to calculate its dual, the group potential profitability HMPI. As input and output prices are not always available to researchers, this motivates the need for a decomposition that does not require such information, i.e., this primal decomposition (see below for how price independent weights can be calculated for the group HMPI). Again, this decomposition reveals the source of changes in group potential productivity—from changes in group productivity without allowing full reallocation (54), or from changes due to allowing full reallocation (58), similarly to the group potential profitability HMPI.

Finally, we can determine a full decomposition of the group potential profitability HMPI, analogous to the efficiency decomposition (47), which we summarize next.

Corollary 3 *We have*

$$\mathrm{PHM}^{*}_{st} = \overline{\mathrm{HM}}_{st} \times \overline{\mathrm{AHM}}_{st} \times \mathrm{RHM}^{*}_{st} \times \mathrm{ARHM}^{*}_{st}, \tag{63}$$

for any periods s *and* t.

The proof of this corollary is via substituting the decompositions of the group and group reallocative profitability HMPIs, (53) and (57) obtained in Propositions (2) and (3), into the decomposition of the group potential profitability HMPI, (60), obtained due to Corollary (2).

A value of (63) is in providing a more complete decomposition of the group potential profitability HMPI, identifying the change due to the group technical HMPI $\left(\overline{\mathrm{HM}}_{st}\right)$, the group allocative HMPI $\left(\overline{\mathrm{AHM}}_{st}\right)$, the group technical reallocative HMPI $\left(\mathrm{RHM}^{*}_{st}\right)$, and the group allocative reallocative HMPI $\left(\mathrm{ARHM}^{*}_{st}\right)$. This gives researchers a clearer indication of the sources of productivity change in the group.[13]

Note again that all the group HMPI measures can be calculated from the individual efficiency scores, after appropriate aggregation. It should be stressed that the aggregation weights used here for aggregating individual efficiency scores are not ad hoc but are derived from economic theory arguments and consistent with other aggregation results in the literature. Incidentally, note that the weights derived are also intuitive measures of the 'economic importance' of each firm in each orientation—observed revenue shares in output orientation and observed cost shares in input orientation. We certainly do not claim that these are the best possible weights, but the value in using them here is that we know how they are derived and from what assumptions. Moreover, they maintain group decompositions analogous to the individual level, and in this productivity context, the derivation scheme reveals which weights to use for each period.

In general, these group HMPI measures may yield different results from the case when the simple (equally weighted) sample mean is used. For example, consider an industry that is dominated by a few large firms and also contains a number of much smaller firms. If the large firms decline in productivity by 10% each (over a given period) and the small firms improve, also by 10% each, a simple mean may suggest that overall industry productivity has improved, but weighting by the economic importance of the firms will reveal that overall industry productivity has actually declined. This is just one example where the two measures (the [weighted] group HMPI measures derived above and the [equally weighted] simple mean) will yield quite different results. This is not to say that the simple arithmetic mean is useless—rather,

[13]Note that the functional forms of the aggregation results discussed here are either arithmetic or harmonic but never geometric. Färe and Zelenyuk (2005) used an alternative approach (based on solving functional equations) to justify a geometric aggregation, which can be advocated for due to the efficiency or productivity indices being multiplicative in their nature. However, their approach did not provide the weights of aggregation (which are typically a much more important part of an aggregation)—they were endogenously selected to be the weights derived from the Koopman-type reasoning we used here. Zelenyuk (2006) pointed out that the first-order approximation relationship of geometric, arithmetic, and harmonic aggregations of productivity indices (with the same weights) and investigated their difference via Monte Carlo experiments. He concluded that for moderate variations of growth in a sample, the difference is very small. One shall remember, however, that the arithmetic and harmonic aggregations are usually more robust to outliers, while the geometric aggregation can be very sensitive to very small values and, will in fact, fail when at least one score is zero (unless its weight is set to zero, with a convention that $0^0 = 1$).

it should be used as a complementary statistic to estimate the first moment of the distribution of HMPI. We would argue, however, that it is important to also compare it to an average that accounts for the economic weight of each observation, e.g., the group measures we have derived here.

5 Practical Matters

Here, we discuss two matters related to the practical estimation of aggregate efficiencies and HMPIs in particular, and especially the most challenging part—estimating the group potential measures and calculating price independent weights.

5.1 Estimation of Group Potential Measures

Note that the group potential HMPI measures are not calculated from the individual efficiency scores, but require calculation directly from the group potential technology. However, with the imposition of two additional assumptions we can recover these measures from the individual scores. These two assumptions are in fact very common for many methods in productivity and efficiency analysis, especially in DEA, which appears to be the most popular in practice for computing productivity indices like HMPIs. Another (and similarly popular) approach of estimation is usually referred to as Stochastic Frontier Analysis (SFA), largely due to Aigner et al. (1977) and many developments since then (e.g., for a recent review, see Kumbhakar et al. (2019) and Parmeter and Zelenyuk (2019)). Specifically, we assume the technology set T_τ^k is the same for all organizations for each period (i.e., $T_\tau^k = T$, for all $k = 1, \ldots, K$) and is convex, and then following Li and Ng (1995) and Nesterenko and Zelenyuk (2007), we get:

$$T_\tau^* \equiv \sum_{k=1}^{K} T_\tau^k = KT_\tau \text{ for all } k = 1, \ldots, K, \text{ for all } \tau, \tag{64}$$

and in turn, for any period τ, we also get:

$$L_\tau^*\left(\overline{Y}_\tau\right) = KL_\tau\left(\widetilde{y}_\tau\right), \tag{65}$$

where $\widetilde{y}_\tau \equiv K^{-1}\sum_{k=1}^{K} y_\tau^k$, so $L_\tau\left(\widetilde{y}_\tau\right)$ is the input requirement correspondence of the average organization in the group in period τ. Following this, the input-oriented group potential efficiencies are the same as the

efficiency measures of the average organization in the group; that is, for all $j = 1, \ldots, n$, we have:

$$\mathrm{ITE}^*_\tau\left(\overline{Y}_\tau, \overline{X}_j\right) = \mathrm{ITE}_\tau\left(\widetilde{y}_\tau, \widetilde{x}_j\right), \tag{66}$$

$$\mathrm{CE}^*_\tau\left(\overline{Y}_\tau, \overline{X}_j, w_j\right) = \mathrm{CE}_\tau\left(\widetilde{y}_\tau, \widetilde{x}_j, w_j\right), \tag{67}$$

$$\begin{aligned}\mathrm{IAE}^*_\tau\left(\overline{Y}_\tau, \overline{X}_j, w_j\right) &= \mathrm{IAE}_\tau\left(\widetilde{y}_\tau, \widetilde{x}_j, w_j\right) \\ &= \mathrm{CE}_\tau\left(\widetilde{y}_\tau, \widetilde{x}_j, w_j\right)/\mathrm{ITE}_\tau\left(\widetilde{y}_\tau, \widetilde{x}_j\right),\end{aligned} \tag{68}$$

where $\widetilde{x}_j \equiv K^{-1}\sum_{k=1}^{K} x_j^k$ for any period j and where ITE, CE, and IAE are as defined in (4), (8), and (10), respectively, with superscript k dropped.

It is worth noting that measure (66) is the aggregate efficiency measure suggested by Førsund and Hjalmarsson (1979). Calculating the group potential measures as the average organization (when all organizations have the same convex technology) enables further intuition. If all organizations were individually efficient but spread across different points of the frontier, the average organization would be inefficient relative to that frontier (i.e., the group potential measure would be inefficient). This is because, though the organizations are individually efficient, if they pooled their resources and technology, they could do better—and the gap between their individually efficient and collectively efficient level is the group reallocative efficiency.

5.2 Price Independent Weights

For practical applications, price information is sometimes unavailable to a researcher (whether input prices, output prices, or both). Shadow prices could be used instead to calculate this aggregation scheme (see Li and Ng 1995). Another approach is to use price independent weights. Färe and Zelenyuk (2003) developed such weights originally, and Färe and Zelenyuk (2007), and Simar and Zelenyuk (2007) extended them, all for the output orientation. Here, we present analogous results for the input orientation.

First, we assume that industry cost share of each input is a known constant (which can vary across time); with these price independent weights can be calculated. Specifically, for a given period τ assume:

$$\frac{w_{n,\tau}\overline{X}_{n,\tau}}{\sum_{n=1}^{N} w_{n,\tau}\overline{X}_{n,\tau}} = b_{n,\tau}, \quad n = 1, \ldots, N, \tag{69}$$

where $\overline{X}_{n,\tau} = \sum_{k=1}^{K} x_{n,\tau}^{k}$ and $b_{n,\tau} \in [0,1]\,(n = 1, \ldots, N)$ are constants (estimated or assumed) such that $\sum_{n=1}^{N} b_{n,\tau} = 1$. With these constants, let

$$\omega_{n,\tau}^{k} = x_{n,\tau}^{k} / \overline{X}_{n,\tau}, \tag{70}$$

be organization k's industry share of the nth input. The input-oriented price independent weights for each organization will then become:

$$W_{\tau}^{k} = \sum_{n=1}^{N} b_{n,\tau} \omega_{n,\tau}^{k}, \quad k = 1, \ldots, K, \tag{71}$$

i.e., a weighted sum comprising the industry input share of an organization for each input, weighted by the industry cost share of the same input, in period τ. Where the $b_{n,\tau}$ cannot be determined, as a special case they can be assumed the same for every input, resulting in an unweighted arithmetic average of input shares in (71) (see Färe and Zelenyuk 2003).

6 Conclusion

This work has summarized various results on aggregation of individual efficiency measures and individual productivity indices into their group analogues. In our discussion, we mainly focused on the Farrell-type efficiency context and on Hicks-Moorsteen productivity indices. We discussed aggregation schemes for these indices with and without allowing full reallocation of inputs and outputs among organizations in the group. This is a valuable extension because of the increasing popularity of these indices, given their appealing theoretical properties and intuitive notion of productivity. The aggregation scheme is theoretically justified, consistent with previous aggregation results, and maintains aggregate decompositions that are analogous to the decompositions at the individual level.

While we have discussed the theoretical measures for aggregate efficiency measures and productivity indices, in practice we only have their estimates. Important extensions therefore will include further development of a bootstrapping methodology. In the DEA context, ideas from Simar and Wilson (1999) and Daskovska et al. (2010) could be merged with those of Simar and Zelenyuk (2007, 2018) and Simar and Wilson (2011). Determining how to estimate group potential technology without requiring the assumptions of convex and identical technology across organizations is a further natural extension.

Acknowledgements The authors thank the editors and anonymous referee, colleagues, and other researchers (especially K. Lovell, E. Diewert, A. Peyrache, C. O'Donnell, P. Rao, Bao Hoang Nguyen, etc.) for their feedback—it helped us improve our paper substantially. Valentin Zelenyuk acknowledges support from ARC Grants (DP130101022 and FT170100401). We remain solely responsible for the views expressed in this article.

References

Aigner, D., C. Lovell, and P. Schmidt. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6 (1): 21–37.

Badunenko, O., D.J. Henderson, and V. Zelenyuk. 2017. The productivity of nations. In *The Oxford handbook of productivity*, chapter 24, ed. E. Grifell-Tatjé, C.A.K. Lovell, and R.C. Sickles, 781–815. New York, NY: Oxford University Press.

Bjurek, H. 1996. The Malmquist total factor productivity index. *The Scandinavian Journal of Economics* 98 (2): 303–313.

Blackorby, C., and R.R. Russell. 1999. Aggregation of efficiency indices. *Journal of Productivity Analysis* 12 (1): 5–20.

Briec, W., and K. Kerstens. 2004, May. A Luenberger–Hicks–Moorsteen productivity indicator: Its relation to the Hicks–Moorsteen productivity index and the Luenberger productivity indicator. *Economic Theory* 23 (4), 925–939.

Briec, W., B. Dervaux, and H. Leleu. 2003. Aggregation of directional distance functions and industrial efficiency. *Journal of Economics* 79 (3): 237–261.

Caves, D.W., L.R. Christensen, and W.E. Diewert. 1982. The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica* 50 (6): 1393–1414.

Charnes, A., W. Cooper, and E. Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2 (6): 429–444.

Daskovska, A., L. Simar, and S. Van Bellegem. 2010. Forecasting the Malmquist productivity index. *Journal of Productivity Analysis* 33 (2): 97–107.

Diewert, W.E. 1983. The measurement of waste within the production sector of an open economy. *The Scandinavian Journal of Economics* 85 (2): 159–179.

Diewert, W.E. 1985. A dynamic approach to the measurement of waste in an open economy. *Journal of International Economics* 19: 213–240.

Diewert, W.E. 1992. Fisher ideal output, input, and productivity indexes revisited. *Journal of Productivity Analysis* 3 (3): 211–248.

Epure, M., K. Kerstens, and D. Prior. 2011. Technology-based total factor productivity and benchmarking: New proposals and an application. *Omega* 39 (6): 608–619.

Farrell, M.J. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A (General)* 120 (3): 253–290.

Färe, R., and D. Primont. 1995. *Multi-output production and duality: Theory and applications*. New York: Kluwer Academic Publishers.

Färe, R., and V. Zelenyuk. 2003. On aggregate Farrell efficiencies. *European Journal of Operational Research* 146 (3): 615–620.

Fare, R., and V. Zelenyuk. 2005. On Farrell's decomposition and aggregation. *International Journal of Business and Economics* 4 (2): 167–171.

Färe, R., and V. Zelenyuk. 2007. Extending Färe and Zelenyuk (2003). *European Journal of Operational Research* 179 (2): 594–595.

Färe, R., and V. Zelenyuk. 2012. Aggregation of scale elasticities across firms. *Applied Economics Letters* 19 (16): 1593–1597.

Färe, R., S. Grosskopf, and V. Zelenyuk. 2004. Aggregation of cost efficiency: indicators and indexes across firms. *Academia Economic Papers* 32: 395–411.

Färe, R., S. Grosskopf, and V. Zelenyuk. 2008. Aggregation of Nerlovian profit indicator. *Applied Economics Letters* 15 (11): 845–847.

Färe, R., X. He, S.K. Li, and V. Zelenyuk. 2019, forthcoming. Unifying framework for Farrell profit efficiency measurement. *Operations Research*.

Førsund, F.R., and L. Hjalmarsson. 1979. Generalized Farrell measures of efficiency: An application to milk processing in Swedish dairy plants. *The Economic Journal* 89 (354): 294.

Fukuyama, H., and W. Weber. 2008. Profit inefficiency of Japanese securities firms. *Journal of Applied Economics* 11 (2): 281–303.

Koopmans, T.C. 1957. *Three essays on the state of economic analysis*. New York: McGraw-Hill.

Kumbhakar, S., C. Parmeter, and V. Zelenyuk. 2019, forthcoming. Stochastic frontier analysis: Foundations and advances. In *Handbook of production economics*, ed. Ray, Chambers, and Kumbhakar. Springer.

Kuosmanen, T., L. Cherchye, and T. Sipiläinen. 2006. The law of one price in data envelopment analysis: Restricting weight flexibility across firms. *European Journal of Operational Research* 170 (3): 735–757.

Kuosmanen, T., M. Kortelainen, T. Sipiläinen, and L. Cherchye. 2010. Firm and industry level profit efficiency analysis using absolute and uniform shadow prices. *European Journal of Operational Research* 202 (2): 584–594.

Li, S.-K., and Y.C. Ng. 1995. Measuring the productive efficiency of a group of firms. *International Advances in Economic Research* 1 (4): 377–390.

Luenberger, D. G. 1996, October. Welfare from a benefit viewpoint. *Economic Theory* 7 (3): 445–462.

Mayer, A., and V. Zelenyuk. 2014a. Aggregation of Malmquist productivity indexes allowing for reallocation of resources. *European Journal of Operational Research* 238: 774–775.

Mayer, A., and V. Zelenyuk. 2014b. An aggregation paradigm for Hicks–Moorsteen productivity indexes. CEPA Working Papers Series WP01/2014, School of Economics, University of Queensland, Australia.

Nesterenko, V., and V. Zelenyuk. 2007. Measuring potential gains from reallocation of resources. *Journal of Productivity Analysis* 28 (1/2): 107–116.

Oks, E., and M. Sharir. 2006. Minkowski sums of monotone and general simple polygons. *Discrete Computational Geometry* 35 (2): 223–240.

Parmeter, C., and V. Zelenyuk. 2019, forthcoming. Combining the virtues of stochastic frontier and data envelopement analysis. *Operations Research*.

Shephard, R.W. 1953. *Cost and production functions*. Princeton: Princeton University Press.

Shephard, R.W. 1970. *Theory of cost and production functions*. Princeton, NJ: Princeton University Press.

Sickles, R., and V. Zelenyuk. 2019. *Measurement of productivity and efficiency: Theory and practice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781139565981.

Simar, L., and P.W. Wilson. 1999. Estimating and bootstrapping Malmquist indices. *European Journal of Operational Research* 115 (3): 459–471.

Simar, L., and V. Zelenyuk. 2007. Statistical inference for aggregates of Farrell-type efficiencies. *Journal of Applied Econometrics* 22 (7): 1367–1394.

Simar, L., and V. Zelenyuk. 2018. Central limit theorems for aggregate efficiency. *Operations Research* 166 (1): 139–149.

Simar, L., and P.W. Wilson. 2011. Inference by the m out of n bootstrap in nonparametric frontier models. *Journal of Productivity Analysis* 36 (1): 33–53.

ten Raa, T. 2011. Benchmarking and industry performance. *Journal of Productivity Analysis* 36 (3): 285–292.

Ylvinger, S. 2000. Industry performance and structural efficiency measures: Solutions to problems in firm models. *European Journal of Operational Research* 121 (1): 164–174.

Zelenyuk, V. 2006. Aggregation of Malmquist productivity indexes. *European Journal of Operational Research* 174 (2): 1076–1086.

Zelenyuk, V. 2011. Aggregation of economic growth rates and of its sources. *European Journal of Operational Research* 212 (1): 190–198.

Zelenyuk, V. 2015. Aggregation of scale efficiency. *European Journal of Operational Research* 240 (1): 269–277.

Intermediate Inputs and Industry Studies: Input-Output Analysis

Victoria Shestalova

1 Introduction

This chapter explains productivity measurement within input-output (I–O) analysis. We aim to cover the literature on both the techniques of performance analysis and the limitations and issues encountered in practice.

The addition of intermediate production into productivity and efficiency measurements has an important advantage, because this allows for a clean separation of the intermediate and final production, thus removing potential measurement biases that arise in other techniques.

When describing advantages of the use of the I–O approach for total productivity growth (TFP) measurement, Wolff (1994) identified five valuable features: 'First, TFP growth can be measured using gross output and incorporating materials as an input. Second, a growth accounting framework can be developed to measure the effect of final output mix changes on aggregate productivity growth. Third, capital can be treated as a produced means of production. Fourth, measures of composite technical change can overcome difficulties engendered by the substitution of externally provided inputs for those previously provided by the firm. Fifth, TFP growth of secondary

V. Shestalova (✉)
Dutch Healthcare Authority (NZa), Utrecht, The Netherlands
e-mail: vshestalova@nza.nl

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_15

production can be analytically separated from that of primary output.' The ability of I–O analysis to account for intermediate product flows and emission flows is especially valuable in the current economic conditions of increasing trade in intermediate inputs and globalization of production activities and environmental pressures. Therefore, these properties make I–O analytical frameworks extremely suitable for performance measurement in these economic conditions.

While the applications of the I–O methodology in industrial performance studies clearly have its merits, simplifying assumptions used by these techniques, notably the assumption of fixed proportions, pose serious limitations in dynamic contexts. Therefore, in this chapter we will also spend some time to discuss limitations.

In the theoretical exposition, we approach the subject of performance measurement from the interface between I–O analysis and frontier analysis. This is done by means of an I–O-based frontier model that combines the advantages of both techniques. Notably, all the prices are endogenous in this model, and the resulting productivity measures are tractable in terms of efficiency. The methodology outlined in this chapter is suitable for performance measurements within both national and international industrial studies, environmental analysis, and other policy-relevant analyses. This type of analysis can be applied in policy evaluation studies and scenario analyses, such as analyses of the gains of free trade or effects of environmental policies. The theoretical exposition of the model draws from the earlier work by ten Raa (2005, 2008), ten Raa and Shestalova (2015a), and Shestalova (2017).

The chapter proceeds as follows: We start by explaining the merits of the I–O accounting framework in the current economic conditions in Sect. 2, after which we will outline the methodology in detail in Sects. 3 (Traditional Model) and 4 (Frontier Model). Section 5 introduces the performance measures embedded in these models. Section 6 pays attention to data requirements and international databases. Section 7 provides some illustrative examples of empirical applications of I–O models for performance measurement. Section 8 concludes.

2 Renewed Interest to I–O Framework

Before discussing the methodology, let us first highlight the reasons for the renewed interest to I–O models in the last decades. The latter has to do with the changing economic conditions that are characterized by increasing

volumes of international trade in intermediate inputs[1] and growing environmental pressures. These two trends pose significant challenges to performance measurement, because failing to account for intermediate inputs and environmental effects can result in significant biases of performance measures. Several recent papers explore potential biases in productivity accounting, arguing about the need to revisit some productivity estimates due to these two trends, on which we elaborate below.

First of all, *globalization* processes in the world economy increase international trade. Since a large share of international trade consists of intermediate inputs, their proper accounting in economic models, and especially in international trade models, becomes increasingly important (Bems 2014). Even though, nearly fifty years ago Melvin (1969) pointed out the possibility of accounting for intermediate inputs in the context of international trade, traditional value-based models have still often been used in practice. Bems (2014) shows that traditional value-added trade models—ignoring production inputs calibrated on gross-flow trade data—result in mismeasured preference weights and price elasticities. These mismeasurements substantially alter model predictions regarding the relative price response to external rebalancing, in comparison with a (preferred) model that is consistent with gross-flow trade data. While early international trade studies were ascribing the trade volume differences to differing production factors—the phenomenon referred to as Heckscher–Ohlin–Vanek paradigm—more recent studies based on currently available data do not support this hypothesis. Instead, they point toward the increasing role of international trade in intermediate products and their differentiation (e.g., Fisher and Marshall 2016).

The international differentiation dictates the need for a proper valuation of internationally traded intermediate products. The gap in intermediate product prices becomes a reason that firms have being increasingly moved abroad in order to benefit from cheaper intermediate product prices. The problem of using incorrect prices in the value of offshored products has been discussed by Houseman et al. (2011), arguing that failing to account for lower production costs in other countries due to the increased substitution of imported goods for domestic goods results in overstated productivity measurement for the United States. Eldridge and Harper (2010) estimated the bias in multifactor productivity in the United States at about 0.1,

[1]Also known as global value chains fragmentation.

suggesting a framework for estimating the effects of imported intermediate inputs in order to solve this problem.[2]

Secondly, *growing environmental pressures and resource prices* suggest that excluding materials, energy and other intermediate inputs from the production function may be increasingly inappropriate. Baptist and Hepburn (2013) explored the relationship between intermediate input intensity, productivity, and national accounts using a panel dataset of manufacturing subsectors in the United States over 47 years. They found a negative correlation between intermediate input intensity and total factor productivity (TFP) both at the aggregated level and at the firm level. The finding that both firms and sectors that are less intensive in their use of intermediate inputs have higher productivity implies that failing to account for intermediate inputs properly would result in biased productivity measurements. Baptist and Hepburn (2013) suggest that current conventions of measuring productivity in national accounts may overstate the productivity of resource-intensive sectors relative to other sectors. This calls for changing national accounting framework to include material inputs and improving the scope and quality of their measurement in order to facilitate the development of policies toward efficient employment of resources, thus increasing productivity.

After establishing the need to account for intermediates and environment, there is still a variety of techniques to be used. I–O analysis has an advantage of providing the most detailed and consistent accounting tool. Due to this and constant improvements of the databases available, I–O analysis is an important tool for performance measurements in industrial studies when it comes to accounting for international product flows, environmental constraints or resource constraints. Notably, useful decomposition techniques developed within the field of I–O analysis enable its application in *policy evaluations* and *policy scenario analyses*. Baumol and Wolff (1994) identify two areas in which I–O analysis becomes indispensable for the formulation of policies: (i) in situations where inputs as well as outputs enter society's objective function directly, such as in the case of employment or pollution and (ii) for open economies, because macroeconomic policy is largely powerless to influence employment or the use of other inputs. Examples are policies aimed at reducing petroleum use through subsidies for other energy sources, curbing the polluting emissions of production processes, and stimulating employment.

[2]Besides, separating intermediate goods from final goods is important for policy analyses, because of different effects of their tariff rate changes on productivity (e.g., Amity and Konings 2007).

3 Traditional I–O Model

In this section, we introduce the basic methodology for I–O analysis, paying attention to the key analytical assumptions on which it rests and consider several extensions.

3.1 Basic Model: Technical Coefficients and Multipliers

An I–O framework introduced by Leontief in the 1930s provides a formal description of relationships between sectors of the economy. In the basic model considered below, the production sectors of the economy are associated with industries each of which produces one homogeneous product, so that there are n different industries and n respective products.[3]

Let us start with the simplest case of a closed economy. Suppose, the production can be separated in n industries producing homogeneous products, where the production by one industry uses the output of other industries. Then, the gross output vector x goes into intermediate and final uses: $x = Ax + f$ where f stands for final uses and A is a $n \times n$ matrix of coefficients in which each coefficient a_{ij} specifies industry j's production requirement from industry i to produce one unit of output. Hence, the gross output of industry i is decomposed as follows:

$$x_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n + f_i$$

Assuming k primary factor inputs, the factor use can be specified in the same vein: $Bx = g$, where g is a k-dimensional factor use vector and the industrial factor input requirements per unit of output are specified by a $k \times n$-dimensional matrix B. Each cell of this matrix represents a ratio of the corresponding primary factor input employment to the gross output produced by the industry. These primary factor inputs are typically capital and labor, but also resources such as land or environment can be included. In a simple case of only one primary input, for example labor, matrix B is represented by a row vector of labor coefficients, $b_l = (b_{l1}, b_{l2}, \ldots, b_{ln})$.

[3]Note that 'industries' is the term used in National Accounting. Although common for IO practitioners, the term sector should be avoided when referring to industries. Note that, SUIOTs are part of National Accounts. In National Accounts, 'sector' refers to the so-called Institutional Sectors (government, households, corporations, etc.) and they are represented in the 'Sectoral Accounts.'

The basic I–O structure of the economy is represented by the system of two equations: $x = Ax + f$ and $Bx = g$.

Three assumptions are central to the model.[4]

Key assumptions

1. Constant returns to scale
2. The technical coefficients are fixed
3. No capacity constraints

The assumption of constant returns to scale is typical in neoclassical growth accounting. Although it may be violated in practice, in competitive industries firms must operate at the optimal scale of production to survive, because the mechanism of entry and exit yields constant returns to scale at the industry level. This agrees with the findings of empirical studies. For instance, Burnside (1996) and Basu and Fernald (1997) provide robust evidence that the manufacturing industry in the United States displays constant returns to scale. While the assumption of constant returns to scale is reasonable for a stationary economy case, it is important to realize that it may still violate in a nonstationary case, when the economy is on a transition path. Also, some industries are characterized by high scale economies, such as railroads.

The second assumption—of fixed proportions—is more restrictive, as it ignores the possibility of both product and factor substitution, existing even in a short run, which is an important limitation (see, e.g., Christ 1955; Duchin and Steenge 2007). The assumption of fixed coefficients can be rationalized by the idea that the technologies need time to adjust and, therefore, cannot change fast. It is implicitly assumed that the coefficients represent an average technology employed in the industry in a given year. Because technical coefficients may change with changed conditions, the model is well suitable for the case of stable technologies, but is less suitable for applications in more dynamic industry contexts, such as ICT.

The same point holds also for the assumption of fixed proportions in consumer preferences. While convenient and theoretically grounded in

[4]These assumptions are central to the model. In addition to them, the traditional model assumes perfect divisibility, no joint production, the independence of consumption from production, and that only the current input and output flows are important (see, e.g., Christ 1955; Duchin and Steenge 2007). As will be discussed in Sects. 3–4, some important assumptions of the traditional model can be released by extending the model.

empirical applications,[5] assuming fixed proportions of final uses (also treating it as independent from production) remains a strong assumption, as the economy may not need to increase the consumption of some products to the same degree as other products. Therefore, if we assume that one must expand all final demand values in the same proportion, then some components would be expanded too much, leading to some waste of resources. However, this effect is small as long as we stay within a range of small deviations from the observed situation, or use this model to make inferences regarding the direction of competitive pressures in the economy.

Finally, the third assumption—of no capacity constraints—suggests that the economy is not restricted in means to expand production, which is probably problematic in the case of a closed economy. Even if it was an open economy, unless the assumption of 'small open economy' (price taker) can be applied, the effects of capacity constraints should not be disregarded.

Note also that we implicitly assume a perfect divisibility of production factors. However, the production factors may be indivisible in practice, and there may be bottlenecks in these factors, restricting the possibilities of the proportional economy expansion in comparison with the model result.

On the positive side, these key assumptions make the model extremely transparent, supporting a straightforward method for impact assessment. Since the coefficients are assumed to be fixed, irrespectively, the production volume, any increase of output requires a proportional expansion of the inputs used. For example, based on equation $(I - A)x = f$, it is easy to establish that in the absence of constraints on primary inputs, an increase of final uses by Δf would translate in the increase of the gross output by $\Delta x = (I - A)^{-1}\Delta f = \sum_{k=0}^{\infty} A^k \Delta f$. In other words, to facilitate the increase in final uses each industry would need to produce intermediate inputs into other industries, and this would create a loop of consequent increases, thus multiplying the effect of the initial change. Because any other industry having linkages to this industry would also have to adjust its output, the total economic effect exceeds the initial change in spending. The ratio of the total effect to the initial effect is called a multiplier. This multiplier property provides a method for impact assessments.[6] The interest to

[5]As will be discussed in Sect. 5.1, this preference structure allows for performance measurement that is clean from inefficiency that arises because of product misallocations between individual consumers (Diewert 1983; ten Raa 2008).

[6]See Coughlin and Mandelbaum (1991), for more detail on multipliers, including those on consumption and employment. Coughlin and Mandelbaum (1991) stressed the following limitations in connection to multipliers, because of which the results of multipliers need to be interpreted with caution. First, the evaluated effects of multipliers are likely to be transitory, as multipliers do not account for price adjustments and behavioural response to changes. Besides, the total effect can easily be overstated

this type of analysis has recently been renewed in the view of the growing environmental pressures and the creation of more detailed environmental accounts. For example, indirect effects need to be accounted for in the evaluation of carbon footprint (Sect. 7 provides some references on this topic).

3.2 Structure of the I–O Table of an Open Economy

Let us next assume that the economy is open and consider the basic structure of the product flows shown in Table 1. The first n rows of this table decompose the gross output of each industry i, x_i, into intermediate uses, domestic final uses and export ex_i as follows:

$$x_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n + f_i + ex_i,$$

Domestic final uses cover household consumption, government purchases, and private investor purchases. These products may partly be imported, as shown by the vector im_f.

On the input side, next to domestically produced intermediate inputs, production of any industry j involves also imported products and primary factor inputs. Therefore, the production value of industry j equals to the column sum of the respective values:

$$x_j = \left(a_{1j} + a_{2j} + \cdots + a_{nj}\right)x_j + v_j + im_j,$$

due to accounting for household consumption as final demand and ignoring re-spending within this economic system or because of 'double accounting' of intermediate products in the model (because of which the total business activity value exceeds its market value, as pointed out by Stevens and Lahr 1988). Second, the effects of an increased spending could easily be overstated because of ignoring supply constraints that may exist in some industries. These constraints relate to both the availability of additional primary inputs and the possibility of their higher employment at unchanged prices. Such supply constraints are especially important in the period of business cycle peaks and less during recessions. Third, multipliers computed based on a regional table account for feedback effects within the economy in question, but not for feedback effects between economies. These effects may be very important for some particular industries. Fourth, the model assumes fixed relationships, while those may change over time, for example due to technical progress of because of substitution between inputs that are induced by changing prices. This critique is especially important since IO tables are often based on data that are a few years old. Fifth, missing data issues arise at the stage of construction of IO table. These are dealt by means of estimation methods, and therefore, the resulting estimates contain some error. Even the calculation of the initial change in spending that is a starting point of the analysis incurs limitations because of the need to account for taxes and heterogeneity of sectoral output. Finally, there is an issue of estimating regional effects of a firm in a new sector.

Table 1 Basic I–O flow table

	Production				Final uses		Total
	Industry 1	Industry 2	…	Industry n	Domestic final uses	Export	Gross output
Products of industry 1	$a_{11}x_1$	$a_{12}x_2$	…	$a_{1n}x_n$	f_1	ex_1	x_1
Products of industry 2	$a_{21}x_1$	$a_{22}x_2$	…	$a_{2n}x_n$	f_2	ex_2	x_2
	…	…	…	…	…	…	…
Products of industry n	$a_{n1}x_1$	$a_{n2}x_2$	…	$a_{nn}x_n$	f_n	ex_n	x_n
Import 1	im_{11}	im_{12}	…	im_{1n}	im_{1f}		
Import 2	im_{21}	im_{22}	…	im_{2n}	im_{2f}		
	…	…	…	…	…		
Import I	im_{I1}	im_{I2}	…	im_{In}	im_{If}		
Value-added payments to primary inputs	v_1	v_2	…	v_n			
Total value	x_1	x_2	…	x_n			

where v_j denotes the value-added payments to primary inputs, and im_j denotes the value of imported products used by industry j. Nowadays, information on imports is available split by products as an imports matrix. Denoting the elements of this matrix by im_{ij}, where subscript j corresponds to the industry j and i runs over the imported intermediate inputs $i = 1, \ldots I$, the intermediate input value in industry j is decomposed as $im_j = \sum_i^I im_{ij}$. Note also, that imports in the exports columns are usually not null in national IOTs, this reflects the so-called re-exports.

Schematically, these flows of products and payments are summarized in Table 1. The last column corresponds to the sum of all purchases along each row, and the last row represents the sum of all payments along each column.

3.3 Extensions of the Traditional Model

While the key assumptions discussed in Sect. 3.1 are critical to the model, thus setting the limits to the application range of the model, some other restrictive assumptions can be released by model extensions, which we discuss in this section. This concerns: (i) accounting for international trade flows, (ii) environmental constraints, (iii) secondary products, (iv) treatment of households, (v) modeling investment, and (vi) endogenizing prices.

International Trade Flows

This is the most natural extension of the model. Modern I–O analyses are conducted with multi-regional models presenting detailed accounts of international trade flows. We provide some examples of existing I–O databases detailing the international product flows and some studies based on these tables in Sects. 6 and 7.

Environment

Another natural extension concerns the inclusion of environmental flows. Environmental constraints can be incorporated in the model in the same way as other primary inputs; thus, pollution coefficients are introduced in the same manner as other input coefficients. Given that both international trade and environment are the most important extensions, we will return to this topic in Sects. 4–6, discussing both theory and applications in more detail.

Secondary Products Within Supply-Use Framework

One issue with the traditional I–O model concerns the difficulty of accounting for secondary products, because of which a supply-use model may be preferable, which we discuss next. Supply and use tables provide an integrated framework for checking consistency and completeness of national accounts data, representing a balancing framework that reconciles income, expenditure, and production data. In this framework, the production side is represented by the use and supply matrices, U and V. The dimension of a supply or use table is product by industry. Under a commodity technology model, 'industry' stands for technology. Each product is produced by a certain industry, i.e., certain technology.[7] Subtraction of the use from the supply matrix defines the net output table of an economy, $U - V$; the summation of net output across industries defines its net output vector. It is straightforward to modify the theoretical exposition of a supply-use model into an I–O model. The matrix of technical coefficients

[7]This results in a product-by-product model matrix. There exists also an alternative model, the so-called industry-technology model associated with an industry-by-industry matrix. See, e.g., ten Raa (2005) for more detail.

is expressed by the formulae: $A = U\left(V^{\mathrm{T}}\right)^{-1}$. In the traditional one-matrix I–O framework, V is assumed to be a diagonal matrix with gross outputs of each industry on the diagonal. Therefore, under this assumption we obtain: $U - V = (I - A)x$.

Household Consumption

An important inconsistency in the basic model is that the level of consumption is assumed exogenous, while the level of income of the labor is endogenous. When production adjusts, so does the value added. The income change should affect household consumption; however, this is not the case in the basic I–O model. This problem, however, can be repaired by making the labor value-added row and the household consumption column interdependent (Duchin and Steenge 2007).

Investment

Another critique concerns the treatment of investment as exogenous factor. A possible solution to this problem is the inclusion of a capital good sector. Duchin and Steenge (2007) consider a dynamic model, in which each sector purchases capital goods in order to assure adequate capital capacity level for its future production. The simplest version of this model assuming a one-year time lag for all the capital goods has the following representation:

$$x_t = A_t x_t + B_{t+1}\max\left(0, k^*_{t+1} - k_t\right) + f_t$$

Here t and $t + 1$ are two consequent years, k_t is the vector of capital capacity in year t, and k^*_{t+1} is the vector of desired capacity in year $t + 1$, B_{t+1} is the matrix of capacity production coefficients, and the other notations are the same as in the basic model in Sect. 3.2. Vector k^*_{t+1} is projected as a moving average of recent past growth of output, starting with the initial value at some year, that is taken as a starting point. The expression $\max\left(0, k^*_{t+1} - k_t\right)$ represents the desired capacity extension in year t.

Endogenous Product Prices

The model can be extended to endogenize prices. Section 4 will introduce the full general equilibrium model which endogenizes both product and

factor input prices. However, before turning to the fully endogenous case, let us first discuss an intermediate step, in which only product prices are endogenous, while factor prices are exogenous.

Duchin (2005) develops a world trade model with m regions, n products, and k factors of production, which represents a closure of a one-region I–O model for international trade with respect to product international prices. While the product prices are endogenous, the model still relies on exogenous factor prices. Therefore, comparative advantage is assumed to be given by exogenously fixed, region-specific technologies, consumption pattern, factor endowments, and also factor prices. Under the assumption of full mobility of products and factors, the model could be specified as follows:

$$\min \pi_1' B_1 x_1 + \cdots + \pi_m' B_m x_m$$

$$\text{s.t. } (I - A_1)x_1 + \cdots + (I - A_m)x_m \geq f_1 + \cdots + f_m$$

$$B_1 x_1 + \cdots + B_m x_m \leq b_1 + \cdots + b_m$$

$$x_1 \geq 0, \ldots, x_m \geq 0$$

where m is the number of regions; f_i and x_i are n-dimensional vectors of final uses and gross output (subscript i running over m regions), A_i are the respective matrices of technical I–O coefficients; B_i are the matrices of factor input requirement coefficients; b_i are k-dimensional endowment vectors of factor inputs; lastly, notation π_i' denotes the exogenous rows of each region's factor prices.

Since product prices are endogenous, any region in this model will engage in trade only to the point at which its imports without trade are worth at least as much as its exports at no-trade prices. However, since primary factor prices are endogenous, a factor will engage in production as long as there is demand at these prices.

The model was applied to the database of 10 regions, 8 products, and 3 primary production factors[8] (Duchin 2005). The empirical application

[8]The database includes North America, Western Europe, Former Soviet Union, Low-Income Asia, China, Japan, Oil-Rich Middle East, Eastern Europe, Middle-Income Latin America, and Rest-of-World (Africa and Low-Income Latin America). The products considered are coal, oil, gas, electricity, mining, agriculture, manufacturing, and services. The primary production factors are labor, capital, and the land. It is assumed that all production sectors need capital and labor, but the land is specific to agriculture.

compares three cases: no trade, world trade, and one-region world. In moving from no-trade model to the world trade model, the products (including services) become tradeable; and in moving from world trade to the one-region world model, the factor mobility is allowed at factor prices of the destination region. The results show that the introduction of international trade leads to little change in the total output in comparison to the first case, except for the energy sector, where the total energy use falls, the fuel mix shifts toward more coal and less gas. However, moving to the one-region world leads to a large change. The latter model results in the lowest value of the objective function. The production takes place using labor-intensive technologies at low labor costs with the greatest export surpluses earned by Eastern Europe and low-income Asia.

The theoretical assumption of exogenous prices is relevant for small open economies, but may be too restrictive for large open economies, such as several economies in these applications. While the move of production toward cheaper inputs confirms the intuition, the comparative advantages of initially low-wage regions could be easily overstated due to the lack of a price adjustment mechanism for factor prices. Therefore, in the next section, we will turn to the case of fully endogenous prices.

4 Frontier I–O-Based Model

After outlining traditional I–O analysis and some of its extensions in the previous section, we are turning to an I–O-based frontier model, which lays the basis to the performance measures that are central to this chapter. The model is based on fundamentals of the economy and internalizes both all the product prices and all the factor input prices. It will be formulated in terms of supply-use tables, since this framework is more general. In order to stay closer to practical applications, we will use a separate notation for each distinct primary factor input considered: capital, labor, and environment. The model exposition in this section draws from the earlier work by ten Raa (2005, 2008), ten Raa and Shestalova (2015a), and Shestalova (2017).

4.1 Model Setup

The model setup specifies the fundamentals of an economy, the technological and behavioral assumptions used, and the physical and financial constraints that must hold in an economy.

Fundamentals

In supply-use framework, the production side is represented by the use and supply matrices, U and V. Here, we apply supply-use framework, following the model exposition in ten Raa and Shestalova (2015a); however, it is straightforward to modify the exposition for the case of I–O coefficient matrices. In a closed economy setting, the net output must fully satisfy the domestic final demand for this output, denoted by product vector f; while in an open economy, the difference can be compensated by international trade. The difference between net domestic output and domestic final demand constitutes net export from the economy, which will be denoted by product vector z. The observed net export satisfies the condition: $z_0 = (V - U)e - f$, where e is the unity vector, all components of which are equal to one. Throughout this section, we assume that all products (including services) are tradeable and we treat trade as the residual between net domestic output and domestic final demand. The case of non-tradeable products can be incorporated by setting the respective components of z at zero. In particular, the case of a closed economy is characterized by the condition: $z = z_0 = 0$.

The economy produces goods using factor inputs. Typically, these are capital and labor; therefore, we present the model for this case. Later, we will show how also environmental input can be included. The allocation of the total employed amounts of capital and labor across industries is given by the respective employment vectors. For notational convenience, we denote the employment row vectors $k = Ke$ and $l = Le$, where K and L are respective diagonal matrices and e is the unity vector, as before. The factors are not always fully employed in production in the observed situation. To account for this, we introduce the diagonal matrices κ and λ of inverse capital and labor utilization rates.[9] For example, a capital utilization rate of 80% for industry 1 corresponds to a full capacity level that is a factor $\kappa_1 = 1/0.8 = 1.25$ higher than utilized in industry 1. The full amounts of capital and labor that are available in the economy are expressed by vectors $K\kappa e$ and $L\lambda e$. This also includes idle amounts of these factors that are not currently employed in production. Note that if factors are mobile across industries, then the allocation across industries can be suboptimal, and it is sufficient to know only the total amounts and the overall utilization rates of both factors.

[9]While the concept of utilization rates is theoretically very straightforward and useful, the data on utilization rates are scarce in practice.

To summarize, formally an economy is represented by the set $E = (k, l, \kappa, \lambda, V, U, f)$ of capital and labor employment and the respective utilization rate matrices, supply and use tables, and domestic final demand vector. All the components are non-negative, and the inverse utilization rate matrices have diagonal values greater or equal to one.

Technological Assumptions

Having explained the main concepts and notations used, we turn to the model assumptions.

Assumption 1
The technology satisfies free disposal of inputs
Free disposal means that the same amount of output can be produced with more inputs, and it is a typical assumption in defining production frontier.

Assumption 2
The technology satisfies constant returns to scale
Constant returns to scale (CRS) means that if industry j produces the net output vector $(V - U)_{\cdot j}$ by using factor inputs K_j and L_j, then for any non-negative value s_j, the scaled net output vector $(V - U)_{\cdot j} s_j$ can be produced by using factor inputs $K_j s_j$ and $L_j s_j$. The value s_j is called the activity level of the industry, where the observed level corresponds to $s_j = 1$. For example, if s_1 equals 1.1, then industry 1 is operated at a 10% higher level than observed. Taken together, these activity levels form the activity vector of the economy s.

Combining the assumption of free disposal and CRS, we obtain that for any non-negative activity vector $s = (s_1, s_2, \ldots, s_n)$, production of the net output $\sum_{j=1}^{n} (V - U)_{\cdot j} s_j = (V - U)s$ requires at least $\sum_{j=1}^{n} K_j s_j = Ks$ capital and $\sum_{j=1}^{n} L_j s_j = Ls$ labor.

Behavioral Assumptions

Assumption 3
Firms maximize profit
The assumption of the profit-maximizing behavior of producers is common in the economic literature (Varian 2010). Denote the product prices by row vector p, the rental rate of capital by row vector r, and the

wage rate by row vector w. The vector of industry profits is expressed by $p(V - U) - rK - wL$.

Assumption 4
Final demand represents domestic preferences
On the consumption side, it is assumed that domestic preferences are represented by the vector of domestic final demand, f. Therefore, the proportions of final demand vector are fixed, and the economy maximizes welfare by expanding the final demand vector to a higher level cf, where c denotes an expansion coefficient of the economy.

While here we impose it by assumption, this behavior can be generated by assuming Leontief preferences (e.g., Leontief 1966). Ten Raa (2008) provides a micro-foundation to the use of this assumption in the context of efficiency measurement, showing that the optimal outcome thus defined would be Pareto improving under any preference structure.

The important advantage of this assumption is that it is sufficient to know only the aggregate domestic final demand and no detailed information at the level of consumers is needed. The domestic final demand vector contains consumption and investment. The theoretical foundation for the inclusion of investment into final demand to account for the whole stream of future consumption is given by Weitzman (1976). It can be shown for competitive economies that domestic final demand measures the present discounted value of future consumption.

Physical Constraints

The material economy must satisfy feasibility constraints on quantities of products and factor inputs.

Products: Material balance constraints imply that consumption cannot exceed production. This holds in the observed data: $V - U \geq f + z_0$. The constraint must hold in any situation, characterized by expanded demand vector cf and activity vector s: $(V - U)s \geq cf + z$.

Production factors: Factor input constraints restrict the amount of factor inputs available. However, the availability of factors depends on factor mobility across industries and countries. Industry-specific features yield the immobility of some types of labor and capital. Similarly, there may be restrictions on cross-border factor mobility. For instance, ten Raa and

Shestalova (2015a) assume that capital is both industry-specific and country-specific, while labor is mobile between industries and countries.

For the sake of notational simplicity and without loss of generality, the exposition of the base model will focus on the case in which both factor inputs are immobile across industries or countries. Then for any production activity vector s, the factor inputs must satisfy the conditions: $Ks \leq K\kappa e$ and $Ls \leq L\lambda e$. The separate restrictions on each industry and country imply that the wage rates and capital returns may differ across industries and countries.

Any other assumption on factor mobility can be incorporated in the model by modifying the respective condition. For instance, the case in which both factors are mobile between industries (but not between economies) can be incorporated by assuming that only the total amount of each factor in the economy is restricted: $e^{\mathrm{T}}Ks \leq e^{\mathrm{T}}K\kappa e$ and $e^{\mathrm{T}}Ls \leq e^{\mathrm{T}}L\lambda e$, where e^{T} is a row vector all elements of which are equal to one. In such a case, the respective equilibrium factor prices will equalize across industries. Alternatively, if a system of several economies is considered, and capital or labor could also move freely between these economies, then only the total amount of each specific factor in the system would need to be restricted, which is obtained by summing up the amounts of factors over all the economies included in this system: $\sum Ks \leq \sum K\kappa e$ and $\sum Ls \leq \sum L\lambda e$. If this condition is imposed, the equilibrium prices of both factors will equalize across countries. Allowing for more factor mobility increases allocative efficiency in the same fashion as in Amores and ten Raa (2014).

Next to the main production factors (capital and labor), the problem may include other relevant factors, such as natural resources (land or minerals). Furthermore, aggregated factor inputs can be disaggregated by type: For instance, labor can be subdivided into skilled and unskilled labor, in which case there will be a separate constraint per type. An example of the latter can be found in ten Raa and Pan (2005).

Budget Constraint

On the value side, economies must satisfy budget constraints that the expenditures cannot exceed the income, thus restricting the ability to borrow. An open economy derives income from three income sources: capital, labor, and international trade; it spends this income on buying products for final consumption.

Assuming all products (including services) be tradeable and prices of domestic and imported products being equalized, this constraint is expressed by the inequality: $pcf \leq r\kappa Ke + w\lambda Le - pz$. In fact, since there

is no satiation, a strict inequality will never occur in equilibrium, as long as $c > 0$. Thus, the equilibrium will be always characterized by the following income-expenditure condition:

$$pcf = r\kappa Ke + w\lambda Le - pz \tag{1}$$

In a closed economy case, net export is zero, $z = 0$, and the last term drops. To define equilibrium in an open economy setting, international trade models often assume balance of payments between imports and exports in equilibrium, implying that $pz = 0$. In both these cases, the national economy budget is equal to the income derived from production factors $r\kappa Ke + w\lambda Le$. The theoretical assumption of balanced payments does not hold in reality, since most countries run a trade deficit or surplus. This imbalance can be preserved in the model by assuming that the value of the optimal net import vector must be within the current budgets, as reflected by the observed international trade pattern:

$$pz \geq pz_0 = S_0 \tag{2}$$

where S_0 denotes the trade surplus achieved at the observed trade level z_0. Therefore, this constraint represents a budget constraint for the welfare maximization problem of an open economy. Since the country uses trade to maximize its own consumption, the constraint will hold with equality in equilibrium. The exact expression for the condition on international trade depends on the model assumptions. In particular, three cases can be distinguished: (i) a closed economy, (ii) a small open economy that is a price taker in international product markets, and (iii) a large open economy.

The feasibility constraints discussed in this section are most commonly present in practical applications. Therefore, the exposition of the base model includes only these main constraints. However, the model can accommodate other types of constraints. For example, Kagawa (2008) and ten Raa and Shestalova (2015a, b) include undesirable outputs such as greenhouse gas emissions and consider environmental policy constraints. The inclusion of additional constraints reduces the scope for efficiency improvements, which we will discuss later in this chapter.

4.2 Optimization Problem as a Linear Programming Problem

This model setup allows us to formulate the optimization problem for the economy. In the simplest case of a closed economy, this problem is a linear programming problem:

$$\max_{s,c} \{c|cf - (V - U)s \leq 0, Ks \leq K\kappa e, Ls \leq L\lambda e, s > 0\}, \tag{3}$$

with the respective dual problem

$$\min_{p,r,w} \{rK\kappa e + wL\lambda e|rK + wL - p(V - U) \geq 0, p \geq 0, r \geq 0, w \geq 0, pf = 1\}. \tag{4}$$

The primary problem expands final demand (with the expansion factor c) subject to product and factor constraints.[10] Therefore, the optimal outcome includes the optimal activity levels (s) of industries together with supporting optimal (shadow) prices of all the constraints included in this problem. In contrast, the dual problem sets constraints on prices, requiring that all the prices are non-negative, while all the profits are non-positive. The shadow variables for the constraints on profits are the respective activity levels of each industry. The normalization condition on prices ensures that the shadow variable for the normalization rule corresponds to the shadow expansion factor c^*. Therefore, based on the main theorem of linear programming, both primary and dual problems produce exactly the same set of optimal values $(c^*, s^*, p^*, r^*, w^*)$, including the expansion factor, the activity vector, and the supporting prices of products and inputs.

The list of all the constraints of both problems includes simple non-negativity constraints on each variable[11] and five other constraints that impose restrictions either on quantities or on prices. It is easy to see that, in equilibrium, if there is slack in a constraint, there is no non-negativity slack in the associated variable and vice versa. This phenomenon is called complementarity slackness. It says that slacks in constraints on products are complementary to the product prices; similarly, slacks in constraints on prices are complementary to the respective optimal quantities.

In the case of a system of open economies linked by trade (but utilizing the own factor inputs), we would obtain a set of similar model equations for each economy. However, the product prices would need to satisfy also the equality constraints for all the economies. This means that the resulting system would not be a simple linear program anymore. Yet, the same principles of complementarity can also be applied to find the solution in this case, as will be shown in the next section.

[10]The theoretical idea is due to Ginsburg and Waelbroeck (1981, pp. 30–31), who consider the maximization of consumption subject to commodity and factor constraints.

[11]Adding a non-negativity constraint on variable c will not change the outcome, because the optimal c is always positive.

4.3 Equilibrium Conditions

Based on the assumptions and constraints specified in Sect. 4.1, we derive a set of equilibrium conditions, which satisfies the optimization problem considered in Sect. 4.2, or a more general optimization model in the case of an open economy.

First, a combination of the assumption of CRS and the assumption of profit maximization rules out the possibility of positive profits in equilibrium. If profits were strictly positive in equilibrium, then at these prices, firms would be able to increase profits by increasing production. However, that would mean that these prices could not be equilibrium prices. Note, however, that negative profit values may be feasible if the respective industry, say industry i, is not active, $s_i = 0$. Therefore, the assumptions made imply the condition of non-positive profits: $(V - U) - rK - wL \leq 0$, with the following complementarity condition:

$$\begin{aligned} s_i > 0 &\Rightarrow (V - U - rK - wL)_i = 0 \quad \text{for all } i \\ (V - U - rK - wL)_i < 0 &\Rightarrow s_i = 0 \quad \text{for all } i \end{aligned} \tag{5}$$

Second, the material balance constraints imply that consumption may not exceed production. However, since there is no idle production in equilibrium, a positive equilibrium product price would imply that the constraint must bind in equilibrium. If the constraint is not binding, the price must be zero. Therefore, there can be only one of the two possibilities:

$$\begin{aligned} p_j > 0 &\Rightarrow (cf - (V - U)s)_j = 0 \quad \text{for all } j \\ (cf - (V - U)s)_j < 0 &\Rightarrow p_j = 0 \quad \text{for all } j \end{aligned} \tag{6}$$

Third, also a positive factor input price would ensure that there is no slack in inputs, leading to the conditions:

$$r > 0 \Rightarrow K(\kappa e - s) = 0 \quad \text{and} \quad K(\kappa e - s) > 0 \Rightarrow r = 0 \tag{7}$$

$$w > 0 \Rightarrow L(\lambda e - s) = 0 \quad \text{and} \quad L(\lambda e - s) > 0 \Rightarrow w = 0 \tag{8}$$

Finally, as long as the expansion factor is positive, $c > 0$, the budget constraint binds in equilibrium:

$$pcf - r\kappa Ke - w\lambda Le + pz_0 = 0 \tag{9}$$

The last term on the left-hand side of Eq. (9), pz_0, is zero in a closed economy setting, but differs from zero in an open economy setting (if we preserve trade imbalance).

The problem specified by the Eqs. (5)–(9) that includes a complementarity condition on each variable is called a complementarity problem. Introducing the notation $\perp$ for the complementary vectors, the set of equilibrium conditions (5)–(9) can be rewritten as follows:

$$\begin{aligned} &0 \leq s \perp rK + wL - (V - U) \geq 0 \\ &0 \leq p \perp (cf - (V - U)s) \geq 0 \\ &0 \leq r \perp K(\kappa e - s) \geq 0 \\ &0 \leq w \perp L(\lambda e - s) \geq 0 \\ &0 \leq c \perp (pcf - r\kappa Ke - w\lambda Le + pz_0) \geq 0 \end{aligned} \tag{10}$$

The solution is given by the set of values $(c^*, s^*, p^*, r^*, w^*)$, including the expansion factor, the activity vector, and the supporting prices of products and inputs.

Since all the prices in model (10) are determined in relative terms, multiplication of all the prices by any fixed positive value gives again a set of equilibrium prices. Hence, the solution is not unique. Uniqueness can be achieved by imposing a normalization rule for prices that equates the total final demand value at equilibrium prices to any constant value v: $p^*f = v$. A natural choice in empirical applications would be to take $v = ef$, equating the final demand value at equilibrium prices to that at observed prices.

Notice the similarity with computable general equilibrium models (CGE), which also generally solve in relative prices. In CGE models, all the values are expressed in terms of the value of one commodity, called the numeraire good, whose price is fixed.

While model (10) is specified for a single economy, it can be easily extended to a more general case of open economies linked by trade by adding a restriction equalizing international prices. See also Ferris and Pang (1997) for more detail on complementarity problems in other settings.

4.4 General Case: A System of Economies Linked by Trade and Environmental Constraints

Let us consider a system of several economies each of which is characterized by its set $E_i = (k_i, l_i, \kappa_i, \lambda_i, V_i, U_i, f_i)$ of capital and labor employment with the respective utilization rate matrices, supply and use tables, and domestic

final demand vector. Assume for the sake of simplicity that all products (including services) are tradeable between economies and all production factors are immobile between economies.

In addition to labor and capital, each economy employs environmental resources. The employment vector of environmental resources of economy i is denoted by M_i, whose components record the damage to the environment by industry, resulting in the total amount of environmental damage $M_i e$, measured in terms of emission volumes of a particular pollutant. Under the assumption of CRS, the proportional increase of each industry's activity leads to the proportional increase in the use of the environmental resource.

Environmental policies are modeled by means of constraints on the amount of pollutants, in the same vein as constraints on capital and labor. For example, a national pollution cap is given by the constraint $M_i(\mu_i e - s_i) \geq 0$, in which μ_i is the fraction of the observed damage, and s_i is the activity vector.

Denoting the shadow prices of emissions by t_i we obtain the complementarity condition:

$$0 \leq t_i \perp M_i(\mu_i e - s_i) \geq 0 \tag{11}$$

The equilibrium conditions for this system are similar to those given by Eq. (10) and can be expressed as follows:

$$\begin{aligned}
&0 \leq s_i \perp r_i K_i + w_i L_i + t_i M_i - p(V_i - U_i) \geq 0 \\
&0 \leq p \perp (cf_i - (V_i - U_i)s_i) \geq 0 \\
&0 \leq r_i \perp K_i(\kappa_i e - s_i) \geq 0 \\
&0 \leq w_i \perp L_i(\lambda_i e - s_i) \geq 0 \\
&0 \leq t_i \perp M_i(\mu_i e - s_i) \geq 0 \\
&0 \leq c_i \perp c_i pf_i - r_i K_i \kappa_i e - w_i L_i \lambda_i e - t_i M_i \mu_i e + pz_{i,0} \geq 0
\end{aligned} \tag{12}$$

Note that in a system of economies with different marginal values of environmental damage, it would be beneficial to shift production in such a way that the environmental damage would be reallocated from the parts with a high marginal value toward the parts with a low marginal value of damage. Such a reallocation could be achieved by means of an international market for emission rights, equalizing shadow prices of environmental resources in each country. A system with internationally tradeable emission permits generates a more efficient allocation of production, thus enabling a higher total consumption in the system of economies. Under tradeable emission rights, the environmental constraints of the participating countries are pooled together, resulting in the following common environmental constraint on the economic system:

$$0 \leq t \perp \sum\nolimits_i M_i(\mu_i e - s_i) \geq 0 \tag{13}$$

where t denotes the international shadow price of emissions in this system.

5 Performance Measures

In this section, we introduce performance measures associated with our model to measure efficiency and productivity growth.

5.1 Efficiency

Accounting for the links and constraints in an economy, the model outlined above naturally embeds the concept of efficiency in terms of expansion possibilities for final demand.

Definition of efficiency
Efficiency is defined by the inverse expansion factor:

$$E = 1/c^* \tag{14}$$

As an example, if an economy's expansion factor equals 1.25, the economy can expand 25%. This means that the efficiency score is $1/1.25 = 0.8$. In other words, the economy can produce its output using just 80% of its resources.

The inverse expansion factor $1/c^*$ can be interpreted in terms of input contraction. It follows from the expenditure-income identity (1), that $1/c = pf/(r\kappa Ke + w\lambda Le + t\mu Me - pz)$. Therefore, efficiency characterizes the minimal resource utilization rate to produce the observed final demand.

The efficiency measure defined above relates to the coefficient of resource utilization introduced by Debreu (1951) for measuring the 'dead loss' that arises in a non-optimal situation, individual preferences been given. Ten Raa (2008) shows that it is possible to remove the dependence on individual preferences by turning to a more conservative measure that removes inefficiency that arises because of product misallocations between individual consumers. According to Diewert (1983), the latter can be done by assuming Leontief preferences, which is also our assumption in this model. Hence, the efficiency measure introduced above is in fact Debreu-Diewert resource utilization measure.

5.2 Total Factor Productivity (TFP)

Here, we discuss productivity measures that arise in both the traditional model and the frontier model, showing how they can be related to each other. While the traditional model attributes total productivity growth to change in technical coefficients, the frontier approach allows also for changes in efficiency and international trade (in an open economy).

We start with the traditional I–O framework and considering the industry-by-industry specification. According to the national accounting identity, the value of the gross production by industry j can be expressed as follows:

$$p_j x_j = \sum_{i=1}^{n} p_i x_{ij} + \sum_{l=1}^{k} w_l g_{lj}.$$

Here p_j denotes the price of the product of industry j, x_j is gross output, x_{ij} is intermediate requirements to this industry from industry i, w_l is the price of factor l, and g_{lj} is the use of this factor by industry j.

Therefore, the rate of TFP growth of an industry is defined as the Solow residual between the growth rate of the gross output and the growth rate of all the inputs, including both primary and intermediate inputs, using the value share of each input in the gross output value as weight. Introducing the notation SR_j for the Solow residual of industry j, we obtain the following formal definition of industrial TFP growth.

Definition of TFP growth at the industry level
The TFP growth is expressed as a Solow residual between the growth of gross output and the growth of inputs:

$$SR_j = \hat{x}_j - \left[\sum_{i=1}^{n} p_i x_{ij} \hat{x}_{ij} + \sum_{l=1}^{k} w_l g_{lj} \hat{g}_{lj} \right] / (p_j x_j). \tag{15}$$

In terms of technical coefficients of the traditional I–O model, $a_{ij} = x_{ij}/x_j$ *and* $b_{lj} = g_{lj}/x_j$*, the TFP growth is expressed as:*

$$SR_j = - \left[\sum_{i=1}^{n} p_i x_{ij} \hat{a}_{ij} + \sum_{l=1}^{k} w_l g_{lj} \hat{b}_{lj} \right] / (p_j x_j). \tag{16}$$

At the total economy level, the intermediate production enters on both input and output side; therefore, its growth does not contribute to TFP growth, leading to the expression of TFP growth as Solow Residual between the net output growth and the primary input growth.

Definition of TFP growth of an economy

The TFP growth of an economy is defined as the Solow residual between the growth of the net output value and the growth of primary factor inputs, expressed by Divisia index:

$$SR = \sum_{i=1}^{n} \frac{p_i y_i}{py} \hat{y}_i - \sum_{l=1}^{k} \frac{w_l g_l}{wg} \hat{g}_l \tag{17}$$

where $y = (I - A)x$ *is net output of the economy, and* g *is a vector of total factor inputs with components* $g_l = \sum_{j=1}^{n} g_{lj}$*, and* p *and* w *are the respective price vectors of outputs and factor inputs.*

The corresponding expression in terms of technical coefficients can be derived as follows. Combining (15) and (17), the aggregate TFP growth of an economy expresses as a weighted average of all industries' TFP growth rates, weights being equal to industries' gross output value shares in the net output value of the economy. Since the gross output exceeds the net output, the sum of these weights is greater than one. The expression for the aggregate TFP growth is known as the Domar decomposition (Domar 1961):

$$SR = \sum_{j=1}^{n} \frac{p_j x_j}{p(I - A)x} SR_j \tag{18}$$

Combining (16) and (18), we obtain the aggregate TFP growth expression in terms of changes of technical coefficients:

$$SR = \sum_{j=1}^{n} \frac{-\sum_{i=1}^{n} p_i x_{ij} \hat{a}_{ij} - \sum_{l=1}^{k} w_l g_{lj} \hat{b}_{lj}}{p(I - A)x} = \frac{-p(dA)x - w(dB)x}{p(I - A)x}. \tag{19}$$

Traditionally, the observed prices and output levels have been used within the I–O field to compute the TFP growth rates using these expressions. An implicit assumption behind this traditional approach is that the economy is in competitive equilibrium. Competitive equilibrium being assumed, prices of inputs are linked by the relationship: $p_j = \sum_{i=1}^{n} p_i a_{ij} + \sum_{l=1}^{k} w_l b_{lj}$.

Since intermediate goods are produced by the economy, their price changes also contribute to TFP changes. Therefore, Aulin-Ahmavaara (1999) suggested that the model can be extended to account for these indirect effects, which results in the so-called effective rates of TFP growth.

In addition to this extension also other extensions are known in the literature, which treat capital input as produced means of production (Peterson 1979; Wolff 1985) or treat both labor and capital as produced by the

economy (Aulin-Ahmavaara 1999), thus resulting in more comprehensive performance measures.

Another extension of the TFP growth measurement is obtained within the general equilibrium framework presented in Sect. 4 of this chapter. The extended TFP growth measure is consistent with the consumer utility maximization behavior under budget constraint, where TFP growth rate is measured by subtracting the change in endowment from the change in consumption (ten Raa 2012). We discuss this performance measure in more detail below, using the notation from Sect. 4.

Definition of TFP growth in a general equilibrium framework
TFP growth in an economy is defined as Solow residual between the growth in overall final demand and the growth in aggregate inputs of the economy:

$$TFP = \frac{pdf}{pf} - \frac{rd(\kappa Ke) + wd(\lambda Le) + td(\mu Me) - pdz_0}{r\kappa Ke + w\lambda Le + t\mu Me - pz_0} \tag{20}$$

where the notation d denotes the change in the respective variable or expression, and the prices $(p, r, w, t) = (p^*, r^*, w^*, t^*)$ *are shadow prices of products and primary factor inputs resulting from the general equilibrium model.*

Since environment and international trade also contribute to the economy, their contribution is also taken into account in this measure. The environment enters the expression in the same way as other primary inputs, while the treatment of international trade depends on the type of the model used (in particular, international trade is zero in a closed economy; traded product price ratios are exogenous in a small open economy, while endogenous in a large open economy).

It is easy to see that the two aggregate performance measures—*SR* and *TFP*—are equivalent in a closed economy with full employment and no environmental constraints. In a traditional one-matrix I–O framework, the latter assumptions imply that $f = (I - A)x$, the capacity utilization coefficients of capital and labor equal 1, and the net export is zero. Therefore, Eq. (19) simplifies into: $SR = \frac{pd[(I-A)x] - rd(Ke) - wd(Le)}{p(I-A)x}$. Next, we denote a row vector of factor prices by $\tilde{w} = (r, w)$ and use the primary factor requirements to construct the matrix of primary factor technical coefficients as follows:

$$B = \begin{matrix} K_1/x_1 & \dots & K_n/x_n \\ L_1/x_1 & \dots & L_n/x_n \end{matrix}$$

Then, Eq. (19) turns into $TFP = \frac{pd[(I-A)x]-\tilde{w}d(Bx)}{p(I-A)x} = \frac{-p(dA)x-\tilde{w}(dB)x}{p(I-A)x}$, which is equivalent to Eq. (18).

5.3 TFP Growth Decomposition

Utilizing complementarity conditions from Sect. 4, Eq. (20) can be transformed into the following sum of three terms:

$$TFP = \frac{pd((V-U)s) - rd(Ks) - wd(Ls) - td(Ms)}{cpf} - \frac{dc}{c} - \frac{pd(z-z_0)}{cpf} = TC + EC + TT \tag{21}$$

The first term in this decomposition—*technical change, TC*—represents the shift of the technological frontier and therefore is equivalent to the Solow Residual evaluated at the equilibrium shadow prices and the optimal production levels resulting from the model. Similarly to the traditional aggregate Solow residual, this term can be attributed to different industries by means of the Domar decomposition (ten Raa and Shestalova 2011).

The second term is *efficiency change, EC*. Since efficiency is defined as the inverse of expansion factor c, a positive change in c corresponds to a negative change in efficiency.[12] This term is similar to the efficiency change term that arises in the decomposition of Malmquist productivity index used by frontier approaches such as data envelopment analysis (DEA) or stochastic frontier analysis (SFA), the link to which will be discussed in Sect. 5.4.

The last term will be called *terms-of-trade effect, TT*. The wedge between the optimal and the observed net export values reflects the gains of free trade. Therefore, this term captures the contribution of free trade to TFP growth. The assumption that the balance of payments is pegged at the observed level follows that $-pd(z-z_0) = (dp)(z-z_0)$. Therefore, this effect can be ascribed to changes in the terms of trade (see Diewert and Morrison [1986], ten Raa and Mohnen [2001, 2002], and Shestalova [2001] on the effect of international trade on productivity).

[12]Efficiency change can further be decomposed into efficiency sources, such as X-efficiency change and allocative efficiency change (Färe and Grosskopf 1996; ten Raa 2012).

5.4 Link to Other Performance Measurement Approaches

The efficiency concept defined in Sect. 5.1 is similar to that from other frontier methods, such as DEA and SFA. These methods define efficiency in terms of distance functions. See, for example, Färe and Grosskopf (1996) for definitions.[13] Under the assumption of CRS, the efficiency measures in terms of input and output distance functions are equivalent. Also in the I–O model considered here, the efficiency measure can be interpreted in terms of both final demand expansion and input contraction.

Another similarity between different frontier methods is the presence of both technical change and efficiency change in TFP growth decompositions. In DEA and SFA, TFP growth is decomposed into two sources: $TFP = TC + EC$, while the frontier I–O-based methodology accounts also for the contribution due to changes in terms of trade.

More generally, ten Raa and Shestalova (2011) show that the TFP growth measure arising in the frontier I–O-based model is interrelated with the main traditional approaches to TFP, namely Solow's residual analysis, the index number approach, and DEA. In particular, any of these measures can be derived within a general unifying framework for TFP growth measurement; the term of technical change arising in all the approaches is equivalent to the shift of the production frontier.

A conceptual difference between different approaches to TFP growth measurement consists in the treatment of prices and the underlying assumption of optimizing behavior. The traditional index number approach uses observed prices, assuming that they are competitive, so that factors are paid their marginal products, and no inefficiency is allowed.[14] Frontier approaches, such as DEA or SFA, make no behavioral assumption, thus allowing for inefficiency. However, since only the production side is optimized, the shadow prices in these analyses do not guarantee all the other equilibrium conditions. In contrast, a frontier I–O-based framework allows for inefficiency as well as defines truly endogenous shadow prices.

[13]While standard data envelopment analysis does not consider intermediate production, it is possible to extend it to account for the effect of intermediate production prices. See Färe and Grosskopf (1996) for more detail. Yet, as discussed later in this section, DEA shadow prices are not fully endogenous, because DEA does not account for material balances or income-expenditure identity.

[14]Alternatively, adjustments could be made for markups and returns to scale (Diewert and Fox 2008).

6 Data Requirements and International Databases

From the theoretical exposition above, it is already clear that the application of I–O analysis requires the use of I–O (or supply-use) tables, as well as data on primary factor inputs. While initial applications of I–O analyses were mainly within national economies, with more globalization and with more available and harmonized international I–O data, there has been a notable increase in the use of I–O techniques in the international economic and policy literature in the last two decades. Several important international databases have been developed over the years, facilitating multi-economy I–O analyses. We briefly describe these developments below, providing some examples of such databases and the issues associated with their construction.

I–O tables are constructed by statistical offices and specialized organizations (e.g., research institutes or universities) based on the same principles as the system of national accounts in order to achieve the consistency both within the tables and with the other data sources. Cross-country harmonization requires the development of a common industrial classification, coordination of price-concepts and definitions used.

International organizations, such as United Nations, the OECD, and Eurostat, put a lot of coordinating efforts, for example, by issuing the guidelines for the national accounts. In 1968, the United Nations issued guidelines for a new overall framework which forms the basis for today's System of National Accounts, describing the compilation of I–O data in the form of use and make tables. Therefore, the modern system is compiled making distinction between commodities and industries (Duchin and Steenge 2007). The latest release of the international statistical standard for the national accounts by the United Nations Statistical Commission is the System of National Accounts 2008 (SNA 2008).[15,16]

In Europe, a consolidated annual supply-use system and derived I–O tables for the European Union and the euro area were for the first time published by Eurostat in 2011. The consolidated European tables result from the aggregation of national tables and a rebalancing treatment of the

[15]https://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf.

[16]As pointed out by the anonymous referee of this chapter, SNA2008 introduces changes with significant impact on GDP figures (i.e., capitalization of R&D) and on trade flows as recorded in national accounts (i.e., international processing). Not all countries have aligned their practices to these new standards yet; therefore, cross-country/years comparison or data usage becomes trickier.

intra-EU import use totals with the intra-EU export supply totals. Until recent, these tables distinguished 59 industries and 59 product groups, nowadays extended to 64 by 64.[17]

With respect to modeling international flows, four levels of I–O tables can be distinguished, namely *regional, intranational, multiregional and interregional* (see Wixted et al. 2006; Kanemoto and Murray 2013, for the discussion on the terminology). In a *regional* I–O model, a technology matrix for one region is specified along with product inflows in this region from other regions and outflows from this region to other regions. A basic example of a one-region table is the I–O table shown at the beginning of this chapter. In contrast, intra-, inter-, and multiregional tables feature multiple regions, with separate technology matrices for each of these regions. The difference between the latter three model types lies in the level of details on international trade flows. In an *intranational* model, only net (rather than gross) outflows or inflows are specified for each commodity in each region. In a *multiregional model* (MRIO), gross flows of each commodity to and from each region are specified. Finally, an *interregional model* (in particular, an intercountry model, ICIO) displays the industry and the region of production as well as the industry and the region of consumption, thus, allowing for analyses of differentiated product flows between economies.

Below, we describe the most-known I–O databases, namely GTAP, Eora, OECD ICIO, and WIOD. While GTAP is compiled for reference years, the three other databases provide time series of tables, thus providing more detail on the development over the period covered.

The most common data source of intercountry CGE modelers is the Global Trade Analysis Project (GTAP) Data Base, which contains complete bilateral trade information, transport, and protection linkages.[18] The GTAP Data Base represents the world economy. The last release—GTAP 9—features 140 regions, 57 commodities for three reference years: 2004, 2007, and 2011. The GTAP breakdown is especially focused on agricultural and energy products (being its strength and the fields where it is widely used, such as agricultural or energy focused studies, or environmental studies in which energy is a key issue).

The Eora MRIO database provides a time series of very detailed IO tables over 1990-2012 with matching environmental and social satellite accounts.

[17] http://ec.europa.eu/eurostat/statistics-explained.

[18] https://www.gtap.agecon.purdue.edu.

The database traces the bilateral flows between 15,909 sectors in 187 countries (Lenzen et al. 2012, 2013).[19]

Another data source is the OECD ICIO database including the tables for 34 industries and 71 countries (regions). The latest release provides a time series 1995–2011.[20] Wood et al. (2019) provides more detail on environmental and labor accounts for these tables. Also, EU ICIOs will be integrated into OECD ICIOs. The production of experimental EU-Inter Country Supply, Use and Input-Output Tables (EU-IC-SUIOTs) falls under the project 'Full International and Global Accounts for Research in Input-Output Analysis' (FIGARO).[21]

The recently developed World Input-Output Database (WIOD[22]) provides new information regarding production fragmentation trends. The newest release—WIOD 2016—represents a series of tables, covering 28 EU countries and 15 other major countries in the world for the period from 2000 to 2014 (Dietzenbacher et al. 2013; Timmer et al. 2014a, 2015).

To facilitate I–O analyses in the environmental field, I–O tables are linked to environmental accounts. The I–O framework for economic and environmental accounts traces environmental flows(Leontief 1970). By linking intermediate inputs to environmental data for all sectors in an economy, an environmentally extended I–O framework allows for an allocation of emission flows to final consumption. Since many goods are internationally traded, this process requires the inclusion of emissions released in other countries to be included.

For example, several large projects were undertaken in Europe to facilitate the estimation of environmental impacts and resource consumption allocations within the EU. Eurostat (2011) has launched several projects to extend individual country supply and use tables (SUTs) with emission data and to create consolidated EU27 tables.[23] The EXIOPOL project founded by the European Commission has integrated research efforts of institutions from different member states to set up a detailed EXIOBASE, currently available for the 2004 and 2007 reference years[24] (Tukker et al. 2009, 2013; Tukker and Dietzenbacher 2013). EXIOBASE2 is one of the most extensive

[19] http://worldmrio.com.

[20] http://www.oecd.org/sti/ind/inter-country-input-output-tables.htm.

[21] http://ec.europa.eu/eurostat/web/economic-globalisation/globalisation-macroeconomic-statistics/multi-country-supply-use-and-input-output-tables/figaro.

[22] http://www.wiod.org.

[23] See also the Eurostat Environmental Accounts. http://ec.europa.eu/eurostat/web/environment/overview.

[24] http://www.exiobase.eu.

environmentally extended MRIO tables available worldwide, which includes also detailed information on water. Furthermore, the WIOD database mentioned above includes links to detailed socioeconomic and environmental satellite accounts (Genty et al. 2012 and Corsatea et al. 2019).[25]

Also, a new database has recently been created for the comparison of the results based on output from the five databases described above (Eora, EXIOBASE, GTAP, OECD ICIOs, and WIOD).[26]

In addition to I–O data, performance analysis requires the data on primary inputs. A well-known international database that includes output and intermediate input data, as well as primary input data regarding employment and aggregate investment[27] by industry in the OECD countries is the OECD Structural Analysis Database (STAN) and also its predecessor—the International Sectoral Database (ISDB). In addition to this database, to generate comparative productivity trends in the EU, the EU KLEMS Growth and Productivity Accounts have recently been created. The EU KLEMS database provides output and input measures on country-industry level, as well as output, input, and productivity growth measures for 25 individual member states Japan and the United States, for the period from 1970 onward.[28] The database distinguishes various capital and labor types, as well as three types of intermediate inputs: materials, energy, and service inputs, which data are derived from supply and use tables of the national accounts. With respect to capital and labor, the database provides a breakdown of capital into ICT and non-ICT assets; it provides a breakdown of hours worked per worker skill type (see O'Mahony and Timmer [2009] for more detail on this database).

While a lot of work has been done to improve the international databases, some practical issues with their construction still pose challenges. In particular, there are difficulties with defining the common industry classification, data confidentiality, accounting for international trade in services and environmental flows (Wixted et al. 2006). These difficulties result in measurement issues, because of which the I–O tables relay on estimates,

[25]Both EXIOBASE and WIOD have SEA (socioeconomic accounts). Note that labor is split in skills preferably based on occupations in EXIOBASE, while in education in WIOD.

[26]For an overview, see http://www.environmentalfootprints.org/mriohome.

[27]For a limited number of countries, also capital stock data are available. Note that capital input series are generally not available from the National Accounts, where only aggregate investment is mostly available.

[28]The maximum disaggregation level includes 72 industries; some variables are at 32 industry level.

which may be subject to errors. The use of stochastic methods helps to overcome some data issues (see, e.g., ten Raa and Steel [1994], ten Raa and Rueda-Cantuche [2007], Rueda-Cantuche and Amores [2010], and Rueda-Cantuche et al. [2013] for more detail on the use of stochastic methods in I–O analysis).

7 Empirical Applications

In this section, we focus on three application areas of I–O framework, namely international and interregional trade, environmental analyses, and policy scenario analyses. In the literature, I–O approach has been applied both to describe the observed allocations and to derive potential efficient allocations. Since both types of analyses are valuable, we include both types of examples in this section, illustrating the applicability of I–O models in these research areas.

7.1 International and Interregional Trade

This section provides examples of I–O models measuring the direction of technical change, TFP growth and efficiency in open economies.

Direction of Technical Change

Since the early 1990s, the trade content is more and more dominated by intermediate goods (e.g., Arto et al. 2015). This trend is especially strong in manufacturing, where the foreign value-added content of production increases on a global scale. By accounting for intermediate production, I–O analysis offers a useful decomposition technique, providing insight in the direction of technical change in different countries.

Timmer et al. (2014b) use the WIOD database, described in Sect. 6, to trace the value added to all capital and labor inputs employed in the production process of final manufacturing goods. To adequately account for differences in labor, they distinguish three levels of worker skills, based on educational attainments. The cells of I–O tables show the origin of all value added needed for the production of a final good, allowing for the representation of the total final output value of each country and industry as a value chain, thus tracing all intermediate good values to their origin. By including 40 countries (representing more than 85% of the global GDP) over the

period 1995–2008[29] and decomposing the final output value of manufacturing by its international content, the study established four major stylized facts: (i) a rapid increase of the foreign share in the value added of a final product, also known as 'international fragmentation'[30]; (ii) a strong shift toward value being added by capital and high-skilled labor, and away from less-skilled labor; (iii) within value chains, an increased high-skilled labor share of advanced nations; and (iv) perhaps surprisingly, an increased capital share of the emerging economies, while a declining share of low-skilled labor in their value added.

TFP Growth Measurement

As explained, I–O analysis offers a framework for the measurement of TFP growth. Here, we focus on applications in the context of frontier models (see Wolff [1994] for more detail on productivity measurement within the traditional I–O framework).

An important advantage of the frontier I–O-based approach to the TFP growth measurement is that it enables decomposition into technical change, efficiency change, and the terms-of-trade effect. Therefore, the model can deliver new insights into the productivity drivers in the context of open economies. For example, ten Raa and Mohnen (2002) analyze productivity changes in the Canadian economy. The Canadian economy is modeled as a small open economy,[31] which is a price taker in the world market; hence, the traded goods prices are exogenously determined, and the exchange rate adjusts producing the terms-of-trade effect on TFP growth. The analysis of the TFP growth reveals the change in the sources of TFP growth over the period 1962–1991. In particular, in the beginning of the period, TFP growth was driven by technical change, then by efficiency, and at the end by the exceptionally high terms-of-trade effect. Similarly, Shestalova (2001) identifies the sources of TFP growth in a model of three large economies

[29]The end of the period was chosen in such a way that the period ends just before the large world economy crisis.

[30]Two important trends in manufacturing are offshoring and international fragmentation. Here, offshoring refers to the transfer of parts of the production process (goods or services) to international locations, while international fragmentation refers to the transformation of previously continuous production processes taking place in one location into those involving sub-processes taking places in different locations.

[31]The model distinguishes 50 industries and 94 commodities, some of which are assumed to be non-tradeable, and distinguishes three capital types: buildings, equipment, and infrastructure.

linked by trade—the United States, Japan, and Europe—in 1985 and 1990. Since this paper considers three large economies, all prices are endogenous. The analysis identifies Japan as having the highest TFP growth in this period, which was driven by technical change.

Efficiency Analysis

The efficiency measure that arises in frontier I–O models provides useful insights on the effects of trade in open economies. Several studies apply frontier-based models for the analysis of efficiency and gains of free trade. In particular, ten Raa and Mohnen (2001) for the analysis of trade between the European and Canadian economies and Sikdar et al. (2005) for measuring the effects of freeing bilateral trade between India and Bangladesh.

The efficiency measure has also been applied within the field of economics of development, focusing on inter-regional flows and consequences of liberalization for welfare allocation between different income groups and inequality. Ten Raa and Pan (2005) and ten Raa and Sahoo (2007) utilize this methodology to study the effects of competitive pressures in the China and India (respectively). Both analyses include detailed modeling of different household groups in order to study welfare allocation between these groups, thus identifying the groups that would be winners and losers from competition.

7.2 Environmental Analysis

There is a trade-off between productivity and environmental objectives, because production increases are often accompanied by emission growth and require more non-renewable resources.[32] Exploring this interrelatedness, input-output approaches have been extended to include environment. Extended I–O models have been applied in different policy contexts, in particular, in CGE environmental analyses (such as GEM-E3[33]), measurements of the effects of environmental regulation on productivity, providing useful insights for international agreements (such as Paris Agreements), or carbon

[32]Many research and policy efforts have been put to internalize the environmental externalities of the economic activity, in particular, in the efficiency realm (see, e.g., Amores and Contreras 2009).

[33]GEM-E3 stands for the General Equilibrium Model for Economy-Energy-Environment. See more detail on https://ec.europa.eu/jrc/en/gem-e3/model.

price information. Below, we consider the use of the integral accounting framework for assigning responsibility for internationally traded greenhouse gas emissions[34] and evaluating pollution haven hypothesis.

While the model introduced in Sect. 4 considers externalities that arise in the production phase, one should keep in mind that externalities may arise also during the consumption or disposal phases, that is during the whole life-cycle of the products, which gives rise to an important field linking I–O and life-cycle analyses.

Because of the capability to trace product flows in a consistent way, I–O analysis provides a natural tool for attributing direct and indirect emissions associated with final consumption to their sources. Multi-region input-output (MRIO) models have increasingly been used for this type of research because they provide an appropriate methodological framework for attributing emissions both at the national and at the supra-national levels. More availability of environmental accounts and higher computer processing capacities have spurred more application of these approaches (e.g., Minx et al. 2009; Wiedmann 2009). Presenting the experience with the use of environmentally extended MRIO models from the UK, Wiedmann and Barrett (2013) argue that these models deliver specific, policy-relevant information that would be impossible to obtain otherwise.

An important methodological advance in this area is the development of integral accounting framework. While the traditional territorial emission accounting allocates emissions to places where they are generated (territorial or producer responsibility), the role of demand for products and international product flows has also been recognized in the literature, giving rise to a more integral emission accounting framework. On the one hand, consumer responsibility approach follows the flows of goods and services, accounting for the embodied emissions. On the other hand, income responsibility considers the flow in the opposite direction of payments for goods and services, thus extending the definition of downstream responsibility beyond the product use and disposal (enabled emissions). Total responsibility reconciles the two (Rodrigues et al. 2006, 2010; Lenzen and Murray 2010; Marques et al. 2012; see also Domingos 2015, for an overview).

Next to these studies, I–O analysis was also applied to evaluate the pollution haven hypothesis. The hypothesis suggests that the implementation of stringent environmental policies by industrialized countries may lead to the relocation of dirty production to developing countries, which typically

[34]It is most commonly defined as the direct and indirect greenhouse gas emissions, measured in tonnes of carbon dioxide equivalent using a 100-year horizon (Fuglestvedt et al. 2003; Minx et al. 2009).

adopt less stringent environmental regulations, thus turning them into pollution havens (Coperland and Tailor 2003). Using traditional I–O methodology, Dietzenbacher and Mukhopadhyay (2007) address the question whether India becomes a pollution haven. Kagawa (2008) uses a frontier (linear programming) model to analyze whether competitive pressure and Japan's compliance with the Kyoto protocol may turn China into a pollution haven. Both papers do not find this effect.

Ten Raa and Shestalova (2015a, b) formulate a complementarity model with environmental constraints, demonstrating the valuable capabilities of the I–O methodology in facilitating policy scenario analyses, which is the topic of the next section.

7.3 Policy Scenario Analysis for International Trade and Environment

Because of its ability to account for both trade in intermediates and environmental externalities, the I–O methodology is useful in evaluating the impact of different policies in open economies which are subject to environmental and other constraints.

Baumol and Wolff (1994) stress the usefulness of I–O analysis in the field of policy analysis, providing examples of such analyses for policies aimed at reducing petroleum use through subsidies for other energy sources, reducing polluting emissions of production processes, and increasing employment.

Policy scenario analyses often need to cope with uncertainty; therefore, any insight into the sensitivity of the model prediction to uncertainty about the model parameters would be very valuable. Uncertainty is especially a concern in the case of environmental policies, where it arises with respect to both the measurement of emissions and the effects of abatement policies. Wiedmann (2009) reviews analyses dealing with uncertainties associated with MRIO models.

Ten Raa and Shestalova (2015a, b) illustrate how the I–O model considered in Sect. 4 of this chapter could be amended to cope with uncertainty in a policy scenario analysis on the effect of pollution caps. The paper considers a system of open economies, linked by trade, and applies the 'economic system' version of the I–O model expressed by Eq. (12), simulating the effect of emission caps on greenhouse gas emissions. Different policy scenarios are modeled by modifying the emission constraints in this system; uncertainty with respect to the underachievement of pollution caps is incorporated by the inclusion of an extra randomly distributed pollution cap parameter for each economy.

8 Concluding Remark

In this chapter, we reviewed the literature on the use of I–O analysis in efficiency and productivity measurements covering also extensions to environmental analysis, explaining the technique, its place in the productivity literature, and its practical applications. We also reviewed the limitations and challenges that play a role in practice, thus defining the application scope of this methodology.

Acknowledgements I am grateful to an anonymous referee for helpful comments and suggestions that contributed to improving this chapter.

References

Amity, M., and J. Konings. 2007. Trade liberalization, intermediate inputs, and productivity: Evidence from Indonesia. *The American Economic Review* 97 (5): 1611–1638.

Amores, A., and I. Contreras. 2009. New approach for the assignment of new European agricultural subsidies using scores from data envelopment analysis: Application to olive-growing farms in Andalusia (Spain). *European Journal of Operational Research* 193 (3): 718–729.

Amores, A., and T. ten Raa. 2014. Firm efficiency, industry performance and the economy: Three-way decomposition with an application to Andalusia. *Journal of Productivity Analysis* 42 (1): 25–34.

Arto, I., J. Rueda-Cantuche, A.F. Amores, E. Dietzenbacher, N. Sousa, L. Montinari, and A. Markandya. 2015. EU exports to the world: Effects on employment and income. EUR—Scientific and Technical Research Reports. http://trade.ec.europa.eu/doclib/docs/2015/june/tradoc_153503.PDF.

Aulin-Ahmavaara, P. 1999. Effective rates of sectoral productivity change. *Economic Systems Research* 11 (4): 349–364.

Baptist, S., and C. Hepburn. 2013. Intermediate inputs and economic productivity. *Philosophical Transactions of the Royal Economic Society A* 371. https://doi.org/10.1098/rsta.2011.0565.

Baumol, W.J., and E.N. Wolff. 1994. A key role for input-output analysis in policy design. *Regional Science and Urban Economics* 24 (1): 93–113.

Bems, R. 2014. Intermediate inputs, external rebalancing and relative price adjustment. *Journal of International Economics* 94 (2): 248–262.

Basu, S., and J. Fernald. 1997. Returns to scale in U.S. production: Estimates and implications. *Journal of Political Economy* 105: 249–283.

Burnside, C. 1996. Production function regressions, returns to scale, and externalities. *Journal of Monetary Economics* 37: 177–201.

Christ, C.F. 1955. A review of input-output analysis. In *Input-output analysis: An appraisal*, 137–182. Princeton: Princeton University Press.

Coperland, B.R., and M.S. Tailor. 2003. *Trade and environment: Theory and evidence*. Princeton: Princeton University Press.

Corsatea, T.D., S. Lindner, I. Arto, M.V. Román, J.M. Rueda-Cantuche, A. Velázquez Afonso, A.F. Amores, and F. Neuwahl. 2019. World input–output database environmental accounts. Update 2000–2016, European Commission.

Coughlin, C., and T. Mandelbaum. 1991. A consumer's guide to regional economic multipliers. *Federal Reserve Bank of St. Louis Review* 73 (1) (January/February): 19–32.

Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19: 273–292.

Dietzenbacher, E., and K. Mukhopadhyay. 2007. An empirical examination of the pollution haven hypothesis for India: Towards a Green Leontief Paradox? *Environmental & Resource Economics* 36: 427–449.

Dietzenbacher, E., B. Los, R. Stehrer, M. Timmer, and G. de Vries. 2013. The construction of the World Input-Output Tables in the WIOD project. *Economic Systems Research* 25 (1): 71–98.

Diewert, W.E. 1983. The measurement of waste within the production sector of an open economy. *Scandinavian Journal of Economics* 85 (2): 159–179.

Diewert, W.E., and K.J. Fox. 2008. On the estimation of returns to scale, technical progress and monopolistic markups. *Journal of Econometrics* 145 (1–2): 174–193.

Diewert, W.E., and C. Morrison. 1986. Adjusting output and productivity indexes for changes in terms of trade. *Economic Journal* 96: 659–679.

Domar, E.D. 1961. On the measurement of technological change. *Economic Journal* 71 (284): 709–729.

Domingos, T. 2015. Accounting for carbon responsibility: The consumer and income perspectives and their reconciliation. *IIOA Newsletter*, November 2015. https://www.iioa.org/news_and_links/newsletters/Newsletter32Nov15-Final.pdf.

Duchin, F. 2005. A world trade model based on comparative advantage with m regions, n goods, and k factors. *Economic Systems Research* 17 (2): 141–162.

Duchin, F., and A. Steenge. 2007. Mathematical models in input-output economics. Rensselaer working papers in Economics 0703.

Eldridge, L.P., and M.J. Harper. 2010. *Effects of imported intermediate inputs on productivity*. Bureau of Labor Statistics, US: Monthly Labor Review.

Eurostat. 2011. Creating consolidated and aggregated EU27 Supply, Use and Input–Output Tables, adding environmental extensions (air emissions), and conducting Leontief-type modelling to approximate carbon and other 'footprints' of EU27 Consumption for 2000 to 2006, Eurostat, Luxemburg.

Färe, R., and S. Grosskopf. 1996. *Intertemporal production frontiers: With dynamic DEA*. Boston: Kluwer Academic Publishers.

Fisher, E.O., and K.G. Marshall. 2016. Leontief was not right after all. *Journal of Productivity Analysis* 46 (1): 15–24.

Ferris, M.C., and J.S. Pang. 1997. Engineering and economic applications of complementarity problems. *SIAM Review* 39: 669–713.

Fuglestvedt, J.S., T.K. Berntsen, O. Godal, S. Sausen, K.P. Shine, and T. Skodvin. 2003. Metrics of climate change: Assessing radiative forcing and emission indices. *Climatic Change* 58: 267–331.

Genty, A., I. Arto, and F. Neuwahl. 2012. Final database of environmental satellite accounts: Technical report on their compilation, WIOD Deliverable 4.6. http://www.wiod.org/publications/source_docs/Environmental_Sources.pdf.

Ginsburg, V.A., and J.L. Waelbroeck. 1981. *Activity analysis and general equilibrium modelling*. Amsterdam: North-Holland.

Houseman, S., C. Kurz, P. Lengermann, and B. Mandel. 2011. Offshoring bias in U.S. manufacturing. *Journal of Economic Perspectives* 25 (2): 111–132.

Kagawa, S. 2008. How does Japanese compliance with the Kyoto Protocol affect environmental productivity in China and Japan? *Structural Change and Economic Dynamics* 19: 173–188.

Kanemoto, K., and Murray, J. 2013. What is MRIO: Benefits and limitations. In *The sustainability practitioner's guide to Multiregional Input-output analysis*, chapter 1, eds. J. Murray and M. Lenzen, 1–9. On Sustainability Book Series. Urbana Champaign, IL, USA: Common Ground Publishing LLC.

Lenzen, M., and J. Murray. 2010. Conceptualising environmental responsibility. *Ecological Economics* 70: 261–270.

Lenzen, M., K. Kanemoto, D. Moran, and A. Geschke. 2012. Mapping the structure of the world economy. *Environmental Science and Technology* 46 (15): 8374–8381.

Lenzen, M., D. Moran, K. Kanemoto, and A. Geschke. 2013. Building Eora: A global multi-regional input-output database at high country and sector resolution. *Economic Systems Research* 25 (1): 20–49.

Leontief, W. 1966. *Input-output economics*. New York: Oxford University Press.

Leontief, W. 1970. Environmental repercussions and the economic structure: An input-output approach. *The Review of Economics and Statistics* 52 (3): 262–271.

Marques, A., J. Rodrigues, M. Lenzen, and T. Domingos. 2012. Income-based environmental responsibility. *Ecological Economics* 84: 57–65.

Melvin, J.R. 1969. Intermediate goods, the production possibility curve, and gains from trade. *The Quarterly Journal of Economics* 83 (1): 141–151.

Minx, J.C., T. Wiedmann, R. Wood, G. Peters, M. Lenzen, A. Owen, K. Scott, J. Barrett, K. Hubacek, Giovanni Baiocchi, A. Paul, E. Dawkins, J. Briggs, Dabo Guan, S. Suh, and F. Ackermann. 2009. Input-output analysis and carbon footprinting: An overview of applications. *Economic Systems Research* 21 (3): 187–216.

O'Mahony, M., and M.P. Timmer. 2009. Output, input and productivity measures at the industry level: The EU KLEMS database. *Economic Journal* 119 (538): F374–F403.

Peterson, W. 1979. Total factor productivity in the UK: A disaggregated analysis. In *The measurement of capital: Theory and practice*, ed. K.D. Patterson and K. Scott, 212–225. London, UK: Macmillan Press.

Rodrigues, J., T. Domingos, S. Giljum, and F. Schneider. 2006. Designing an indicator of environmental responsibility. *Ecological Economics* 59 (3): 256–266.

Rodrigues, J.F., T.M. Domingos, and A.P. Marques. 2010. *Carbon responsibility and embodied emissions: Theory and measurement*. New York: Routledge.

Rueda-Cantuche, J.M., and A.F. Amores. 2010. Consistent and unbiased carbon dioxide emission multipliers: Performance of Danish emission reductions via external trade. *Ecological Economics* 69 (5): 988–998.

Rueda-Cantuche, J.M., E. Dietzenbacher, E. Fernández, and A.F. Amores. 2013. The bias of the multiplier matrix when supply and use tables are stochastic. *Economic Systems Research* 25: 435–448.

Shestalova, V. 2001. General equilibrium analysis of international TFP growth rates. *Economic Systems Research* 13: 391–404.

Shestalova, V. 2017. Efficiency analysis in an input-output based general equilibrium framework. In *Handbook of productivity analysis*, eds. T. ten Raa and E. Wolff. Cheltenham, UK: Edward Elgar Publishing.

Sikdar, C., D. Chakraborty, and T. ten Raa. 2005. A new way to locate comparative advantages of India and Bangladesh on the basis of fundamentals only. In *Essays on international trade, theory and policy for the developing countries*, chapter 6, ed. R. Acharyya, 169–97. Kolkata: Allied Publishers Private Limited.

Stevens, B.H., and M.L. Lahr. 1988. Regional economic multipliers: Definition, measurement, and application. *Economic Development Quarterly* 2, 88–96.

ten Raa, T. 2005. *The economics of input-output analysis*. Cambridge: Cambridge University Press.

ten Raa, T. 2008. Debreu's coefficient of resource utilization, the Solow residual, and TFP: The connection by Leontief's preferences. *Journal of Productivity Analysis* 30: 191–199.

ten Raa, T. 2012. Performance measurement in an I–O framework. *Journal of Economic Structures* 1: 2.

ten Raa, T., and P. Mohnen. 2001. The location of comparative advantages on the basis of fundamentals only. *Economic Systems Research* 13 (1): 93–108.

ten Raa, T., and P. Mohnen. 2002. Neoclassical growth accounting and frontier analysis: A synthesis. *Journal of Productivity Analysis* 18: 111–128.

ten Raa, T., and H. Pan. 2005. Competitive pressures on China: Income inequality and migration. *Regional Science & Urban Economics* 35: 671–699.

ten Raa, T., and A. Sahoo. 2007. Competitive pressure on the Indian households: A general equilibrium approach. *Economic Systems Research* 19 (1): 57–71.

ten Raa, T., and J. Rueda-Cantuche. 2007. Stochastic analysis of input-output multipliers on the basis of use and make matrices. *Review of Income and Wealth* 53 (2): 318–334.

ten Raa, T., and V. Shestalova. 2011. The Solow residual, Domar aggregation, and inefficiency: A synthesis of TFP measures. *Journal of Productivity Analysis* 36 (1): 71–77.

ten Raa, T., and V. Shestalova. 2015a. Supply-use framework for international environmental policy analysis. *Economic Systems Research* 27 (1): 77–94.

ten Raa, T., and V. Shestalova. 2015b. Complementarity in input-output analysis and stochastics. *Economic Systems Research* 27 (1): 95–100.

ten Raa, T., and M.F.J. Steel. 1994. Revised stochastic analysis of an input-output model. *Regional Science and Urban Economics* 24: 361–371.

Timmer, M., E. Dietzenbacher, B. Los, R. Stehrer, and G. de Vries. 2014a. The World Input-Output Database (WIOD): Contents, concepts and applications. GGDC Research Memorandum 144, Groningen Growth and Development Centre, The Netherlands.

Timmer, M., A. Erumban, B. Los, R. Stehrer, and G. de Vries. 2014b. Slicing up global value chains. *The Journal of Economic Perspectives* 28 (2): 99–118.

Timmer, M.P., E. Dietzenbacher, B. Los, R. Stehrer, and G.J. de Vries. 2015. An illustrated user guide to the World Input-Output Database: The case of global automotive production. *Review of International Economics* 23: 575–605.

Tukker, A., E. Poliakov, R. Heijungs, T. Hawkins, F. Neuwahl, J.M. Rueda-Cantuche, S. Giljum, S. Moll, J. Oosterhaven, and M. Bouwmeester. 2009. Towards a global multi-regional environmentally extended input-output database. *Ecological Economics* 68: 1929–1937.

Tukker, A., and E. Dietzenbacher. 2013. Introduction to global multiregional input-output frameworks: An introduction and outlook. *Economic Systems Research* 25: 1–19.

Tukker, A., A. de Koning, R. Wood, T. Hawkins, S. Lutter, J. Acosta, J.M. Rueda Cantuche, M. Bouwmeester, J. Oosterhaven, T. Drosdowski, and J. Kuenen. 2013. EXIOPOL—Development and illustrative analyses of a detailed global MR EE SUT/IOT. *Economic Systems Research* 25: 50–70.

United Nations Statistical Division. 2008. System of National Accounts.

Varian, H.R. 2010. *Intermediate microeconomics: A modern approach*. 8th ed. New York and London: W.W. Norton.

Weitzman, M.L. 1976. On the welfare significance of national product in a dynamic economy. *Quarterly Journal of Economics* 90: 156–162.

Wiedmann, T. 2009. A review of recent multi-region input–output models used for consumption-based emission and resource accounting. *Ecological Economics* 69 (2): 211–222.

Wiedmann, T., and J. Barrett. 2013. Policy-relevant applications of environmentally extended MRIO databases—Experiences from the UK. *Economic Systems Research* 25 (1): 143–156.

Wixted, B., N. Yamano, and C. Webb. 2006. Input-output analysis in an increasingly globalised world: Applications of OECD's harmonised international tables, OECD science. Technology and Industry Working Papers, 2006/07, OECD Publishing.

Wolff, E.N. 1985. Industrial composition, interindustry effects and the US productivity slowdown. *Review of Economics and Statistics* 67: 268–277.

Wolff, E.N. 1994. Productivity measurement within an input-output framework. *Regional Science and Urban Economics* 24 (1): 75–92.

Wood, R., K.S. Wiebe, M. Simas, S.Y. Schmidt, J. Kuenen, T.D. Corsatea, and A.F. Amores. 2019. Environmental and labour accounts for OECD inter-country input-output tables 2010–2013, European Commission.

Modelling Environmental Adjustments of Production Technologies: A Literature Review

Externalities and Environmental Studies

K. Hervé Dakpo and Frederic Ang

1 Introduction

The mainstream definition of neoclassical economics has long focused on the analysis of scarce resources that can be marketed through demand and supply. As underlined in Mehta (2013), the 'scarcity' concept in neoclassical economics concerns mostly resource allocation through markets where there exist infinite substitution possibilities of abundant resources to scarce ones. Yet the non-priced outcomes of production activities have to some extent been overlooked. Following the publication of *Our Common Future* (Keeble 1988), the concept of sustainable development has come to the political forefront. Sustainable production plays a key role to this end. Production processes frequently generate not only the intended outputs,

K. H. Dakpo
Economie Publique AgroParisTech, INRA,
Université Paris-Saclay, Thiverval-Grignon, France
e-mail: k-herve.dakpo@inra.fr

F. Ang (✉)
Business Economics Group, Wageningen University,
Wageningen, The Netherlands
e-mail: frederic.ang@wur.nl

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_16

but also negative and positive environmental externalities. This calls for an extension of the usual production framework.[1] The main challenge for the empirical analyst is to do this in an axiomatically accurate way. This chapter provides a theoretical review and discussion of the existing approaches to deal with negative externalities.[2] This mainly includes models that consider pollution as an input or as an output under the weak disposability assumption (WDA); materials balance models and the multi-equation modelling of pollution-generating technologies. Though the last decades have seen an emergence of such models, the existing reviews have focused on a methodological discussion. For instance, Dakpo et al. (2017b) and Adler and Volta (2016) have discussed pollution-generating technologies regarding the non-parametric framework of data envelopment analysis (DEA), and Zhou et al. (2018) have undertaken a bibliometric analysis of pollution studies. In this chapter, we present a more theoretical discussion without an explicit allusion to performance benchmarking (DEA or stochastic frontier analysis—SFA). Our work is grounded on Førsund (2009, 2017, 2018). This chapter complements these latter papers with a specific discussion of each of the models that deal with bad outputs in a more exhaustive way. We have also provided a brief discussion related to abatement technologies. If traditionally environmental bads are mostly encountered in the literature, environmental goods (positive externalities) must also be credited to decision-making units (DMUs) that produce them. For instance, in livestock farming, carbon sequestration in grasslands area is viewed as one of the most important pillars for mitigating greenhouse gas (GHGs) emissions from this sector and therefore this positive non-marketed output should not be neglected. In this chapter, we propose how to deal with environmental goods. Finally, this chapter summarizes the lessons from the different models discussed as well as the challenges that need to be dealt with in modelling environmentally adjusted production technologies.

[1]Pigou (1920) initiated the integration of externalities into a partial static analysis framework and supported the idea that public intervention is a vector of efficiency. Welfare economics has thus focused on the processes of internalization of these externalities.

[2]In this chapter, we refer to negative externalities as environmental bads, pollution, residuals, detrimental outputs, undesirable outputs, bad outputs, wasters or unintended outputs. On another hand, we refer to the traditional outputs as good outputs or intended outputs. About positive externalities, we refer to them as environmental goods.

2 The Thermodynamics of Pollution Generation

Though the concepts of thermodynamics are largely applied to physics or chemistry, they have been extended to various life sciences (biology, ecology, psychology) and lately to economics. The first discussion of economic systems in relation to thermodynamics and the second law can be dated to Georgescu-Roegen (1971). More discussion on the entropy law and environmental economics can be found in Ayres (1998). The first law has been introduced in economics by Ayres and Kneese (1969) in what is known as the materials balance principle (Kneese et al. 1970). The materials balance principle acknowledges mass conservation and, applied to the economic system, it translates as: 'the mass of all material inputs from the environment to the economy, ignoring flows from the environment directly to the final consumption sector, equals the mass of inputs to the intermediate product sectors; the mass of inputs to the intermediate product sectors equals the mass of products supplied to the final consumption sector plus the mass of residuals discharged to the environment minus the mass of materials recycled; and the mass of all final products equals the mass of materials recycled plus residuals generated by the final consumption sector. Assuming no accumulation or recycling, the mass of all inputs from the environment must equal the mass of all residuals discharged to the environment' (James 1985, pp. 271–272). Considering a circular flow economy, the materials balance can be described as in Fig. 1.

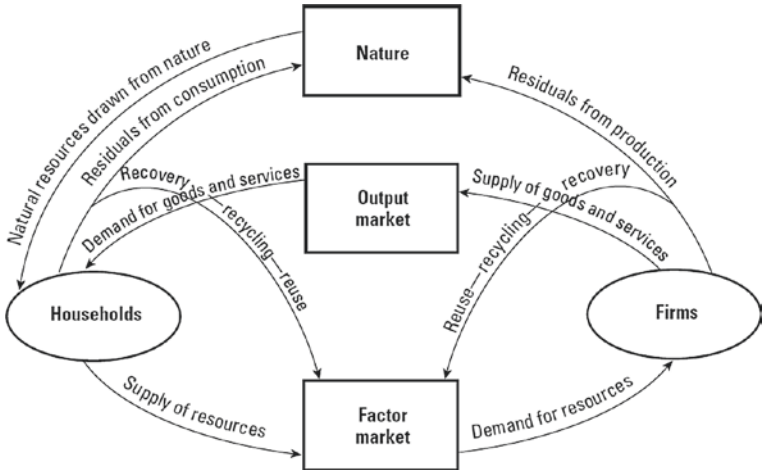

Fig. 1 Materials balance in a circular economy (*Source* Callan and Thomas 2009, p. 5)

It is worth stressing that while Fig. 1 describes the material flows between different economic agents and nature, we will only focus on the firms' side in this chapter as mentioned in the introduction section.

To summarize, the main message on the role of thermodynamics on economic activities has been adequately conveyed in the following saying by Faucheux (1994, p. 8): 'energy and mass conservation, together with the second law of thermodynamics (entropic irreversibility), implies the inevitability of unwanted by-products or waste energy in the course of economic production and consumption'. Baumgärtner et al. (2001) used the term of 'joint production' to describe economic systems that simultaneously produce desirable and undesirable goods.

Formally, the mass conservation equation can be written as follows:

$$x_M = y + z \tag{1}$$

where x_M represents the material inputs, y the desirable outputs and z the residuals also known as the 'uncontrolled byproducts' (Rødseth and Romstad 2013). In (1), x_M, y and z are all expressed in the same mass content. Formula (1) can be generalized to several material inputs and outputs using mass contents as in formula (2).

$$a' x_M = r' y + c' z. \tag{2}$$

In (2), r can be viewed as the vector of material coefficients per unit of mass in the desirable outputs. In many applications (especially in the case of pollution), these coefficients are set to zero (Coelli et al. 2007; Rødseth and Romstad 2013; Hampf and Rødseth 2015). Due to non-homogeneity in the inputs and outputs, and also the role of some external factors (like weather), the material flow coefficients (a, c, r) may differ from one DMU to another.

According to Førsund (2017), *the mass balance equations in* (1) *and* (2) *are accounting identities which hold true for any observation (efficient or not)* and introduce some restrictions on the specification of the production technology. Besides, the mass balance equation as accounting identity does not explicitly explain how residuals are actually generated (Førsund 2009, 2017). As a consequence, we do believe that the mass balance equation cannot be directly used to derive optimal economic behaviours. Finally, the relations in (1) and (2) do not include non-material or service inputs x_S, which can be used to some extent as 'dematerialization' to reduce the levels of residuals by scaling down the amount of material inputs (Ayres and Ayres 2002, p. 9; De Bruyn 2002; Baumgärtner 2004, p. 311). This issue will be discussed in the next sections.

The second law of thermodynamics implies the following (Baumgärtner et al. 2001; Pethig 2006):

$$\frac{\mathrm{d}z}{\mathrm{d}x_M} > 0 \tag{3}$$

The relation in (3) has been included in the modelling of polluting technology by Rødseth (2015) as the axiom of input essentiality for bad output.

In practice, a production process involves a great number of variables among which many are unobservable and therefore not accounted in the mass balance equation. Thereby, the equality in relations (1) and (2) may be violated. Since not all the variables can be accounted for, we consider that what matters is a proper modelling of the production technology without explicit inclusion of mass balance equations.

3 Modelling Pollution-Generating Technologies

In this section, we discuss two ways of representing pollution-generating technologies. The first one is based on the single representation of pollutants, considering pollution as either an input or an output and the second one is the multi-equation representation, which describes the production system as a collection of different subprocesses. Given the argument made above, stating that all physical production activities are governed by the materials balance principle, the discussion that follows will refer to this concept to discuss all models.

3.1 Single Representation of Pollution-Generating Technologies

Frisch (1965, p. 10) has defined a single production technology as follows: 'if a production process is such that it results in a single, technically homogeneous kind of goods or services, we call it single production'. In this framework, residuals have been treated either as inputs or as outputs. Let's denote $x = (x_1, x_2, \ldots, x_K) \in \mathbb{R}_+^K$ the vector of all inputs (material x_M and service inputs x_S), $y = \left(y_1, y_2, \ldots, y_Q\right) \in \mathbb{R}_+^Q$ the vector of good outputs and $z = (z_1, z_2, \ldots, z_R) \in \mathbb{R}_+^R$ the vector of residuals.

Residuals as Inputs

A single representation of the production technology Ψ, where residuals are treated as inputs, is:

$$\Psi = \left\{(x, y, z) \in \mathbb{R}_+^{K+Q+R} | (x, z) \text{ can produce } y\right\} \tag{4}$$

The graph technology in (4) can also be described using input and output sets (correspondences) that summarize the properties of the isoquant curves. The input correspondence is:

$$L : \mathbb{R}_+^Q \rightarrow L(y) = \left\{(x, z) \in \mathbb{R}_+^{K+R} | (x, y, z) \in \Psi\right\} \tag{5}$$

In (5), $L(y)$ is the input set or the input requirement set (Färe and Grosskopf 1996). Similarly, the output correspondence is represented by:

$$P : \mathbb{R}_+^{K+R} \rightarrow P(x, z) = \{y \in \mathbb{R}_+^Q | (x, y, z) \in \Psi\} \tag{6}$$

where $P(x, z)$ is the output set. These representations of the production technology are very helpful in focusing on particular aspects (marginal rates of substitution for example) of production such as substitution among outputs or inputs.

Equivalently, the production technology described in (4) can also be represented using the transformation function:

$$\Psi = \left\{(x, y, z) \in \mathbb{R}_+^{K+Q+R} | F(x, y, z) \leq 0\right\} \tag{7}$$

For any point located at the boundary of the technology or at the transformation frontier, we can write the following:

$$F(x, y, z) = 0, F_y = \frac{\partial F}{\partial y} \geq 0, F_x = \frac{\partial F}{\partial x} \leq 0, F_z = \frac{\partial F}{\partial z} \leq 0 \tag{8}$$

The signs of the derivatives in (8) are related to the monotonicity conditions, which imply that all the variables are strongly disposable. In other words, we have:

$$\begin{aligned} (x, y, z) \in \Psi, \hat{y} \leq y \Rightarrow (x, \hat{y}, z) \in \Psi \\ (x, y, z) \in \Psi, \hat{x} \geq x, \hat{z} \geq z \Rightarrow (\hat{x}, y, \hat{z}) \in \Psi \end{aligned} \tag{9}$$

Moreover, in the presence of inefficiency, the transformation relation in (8) can be written as $F(x, y, z) < 0$. Let's assume $F_z < 0$, then using implicit function theorem, one can express z as a function of x and y (Murty et al. 2012):

$$z = h(x, y), F(x, y, h(x, y)) = 0 \tag{10}$$

In (10), the trade-off between desirable outputs and residuals is captured by the following:

$$\frac{\partial z}{\partial y} = \frac{\partial h(x, y)}{\partial y} = -\frac{F_y}{F_z} \geq 0 \tag{11}$$

This positive correlation between desirable and residuals has been—to our point of view—the sole reason for studies that consider residuals as inputs. For instance, considering that pollution generates social costs and that an input orientation is straightforwardly interpreted in terms of costs savings (minimization), some authors treat pollution as an input in the production technology (Tyteca 1997; Courcelle et al. 1998; Hailu and Veeman 2000; De Koeijer et al. 2002; Reinhard et al. 2002; Prior 2006; Telle and Larsson 2007; Yang and Pollitt 2009; Mahlberg et al. 2011; Mahlberg and Sahoo 2011). Their argument is that detrimental variables are considered as an indirect 'use of natural resources' (Dyckhoff and Allen 2001), or the use of environment as a 'free' input (Paul et al. 2002; Considine and Larson 2006), and that, empirically, pollution and good outputs are generally positively correlated.

The trade-off in (11) is obtained by assuming that the input levels are fixed. Yet under this condition, considering the mass equation in (1) and differentiating it totally implies:

$$0 = \mathrm{d}y + \mathrm{d}z \Leftrightarrow \frac{\mathrm{d}y}{\mathrm{d}z} = -1 < 0 \tag{12}$$

The conservation law implies a negative constant correlation between desirable outputs and residuals for fixed levels of material input, which contradicts the results in (11). Given that the materials balance is not a technology but rather an accounting identity, the correlation in (12) implies that two observations may have the same level of x_M but different values of y and z.

Let us suppose that the material inputs are no longer fixed. Therefore, we have:

$$\mathrm{d}x_M = \mathrm{d}y + \mathrm{d}z \Leftrightarrow \frac{\mathrm{d}y}{\mathrm{d}z} = \frac{\mathrm{d}x_M}{\mathrm{d}z} - 1 \geq 0 \tag{13}$$

The last constraint in (13) is always non-negative because the mass of the material inputs is greater or equal to the mass of the residuals (thereby $\frac{\mathrm{d}x_M}{\mathrm{d}z} \geq 1$). From (13), it appears that the trade-offs in (11) are only feasible if the material inputs are no longer held fixed. In other words, *ceteris paribus*, trade-offs in (11) are not 'thermodynamiaclly' feasible.

Similarly, the trade-offs between residuals and all the inputs (for fixed level of intended outputs) are obtained by:

$$\frac{\partial z}{\partial x} = \frac{\partial h(x,y)}{\partial x} = -\frac{F_x}{F_z} \leq 0 \tag{14}$$

The trade-offs in (14) are physically inconsistent because the relation in (14) is valid for all the inputs, be it materials or non-materials. Yet from the discussion in Sect. 2, material inputs generate residuals as in the expression of the entropy law in formula (3). However, for service inputs, the relation in (14) will simply reflect dematerialization.

It appears that the modelling of residuals as inputs has been mainly guided by the search for a positive relation between desirable outputs and residuals, while all other relations have been overlooked. Yet until now all the demonstrations consistently reject the modelling of pollution as input given the materials balance principle. At this point, one may wonder why many empirical studies have considered residuals as inputs despite the aforementioned issues. We believe that the simplicity of this modelling facilitates empirical illustrations.

The relation in (14) is a consequence of modelling residuals as inputs. It is worth noting that, given the input essentiality axiom, zero residuals imply also zero material inputs and it is clear that production cannot occur with zero level of material inputs.[3] This last feature has been considered restrictively as the null-jointness property in Shephard and Färe (1974) and Färe et al. (1989). The null-jointness of desirable outputs and residuals is formalized as:

$$(x, y, z) \in \Psi \text{ and } z = 0, \text{ then } y = 0 \tag{15}$$

[3]See Pethig (2003) and Baumgärtner (2004) for more discussion on the Inada conditions.

As formulated in (15), the null-jointness property is very restrictive and is incompatible with the materials balance principle. Following the discussion in the previous paragraph, a proper null-jointness property can be written as:

$$(x, y, z) \in \Psi \text{ and } x_M = 0, \text{ then } z = 0 \text{ and } y = 0 \tag{16}$$

Moreover, from the mass balance equation in (1) we know that $y < x_M$, which implies that $F_{x_M} < 1$, which in turn rules out the Inada condition $\lim_{x_M \to 0} F_{x_M} = \infty$.

It is worth noting that the amount of residuals can be reduced through end-of-pipe activities or investment into new technologies. As aforementioned, however, wastes cannot be infinitely recycled. Moreover, as underlined by Førsund (2009), it is difficult to imagine how increasing residuals (all the other inputs fixed) will increase the intended output. Clearly, residuals are consequences of production processes and not the opposite.

The production technology presented in (5) can also be described using the Shephard input distance function (Shephard 1953, 1970) defined as:

$$D_i(x, y, z) = \sup \left\{ \theta : \left(\frac{x}{\theta}, \frac{z}{\theta} \right) \in L(y) \right\} \tag{17}$$

Considering the mass condition in (1) and the distance function, we have:

$$\frac{x_M}{\theta} = y + \frac{z}{\theta} \Leftrightarrow \frac{1}{\theta}(x_M - z) = y \tag{18}$$

As pointed out in Coelli et al. (2007), the only solution to the problem in (18) is when $\theta = 1$, which implies no inefficiency can be measured under the case where residuals are treated as inputs.[4] There seems to be a misuse of the materials balance in demonstration (18), where frontier levels of the variables are inserted in the materials balance identity. This may not be the appropriate way of using the materials balance for an unobserved frontier point. Besides, as said earlier, the materials balance is also valid for inefficient points. Therefore, a point worth discussing relatively to the demonstration in (18), which is also put forward in Coelli et al. (2007), is the disposability of the intended outputs. In the technology described by the input distance function in (17), the intended outputs are strongly disposable. Yet in demonstration (18), those outputs are held fixed. If we introduce this disposability of intended outputs in (18), we have:

[4]See Hoang and Coelli (2011) for the case of directional distance function.

$$\frac{x_M}{\theta} = \hat{y} + \frac{z}{\theta} \Leftrightarrow \frac{1}{\theta}(x_M - z) = \hat{y} \leq y \tag{19}$$

and then some inefficiency can be measured especially if $\hat{y} < y$. The demonstration in (19) complies with the materials balance but it doesn't mean that the model is correct given the trade-off issues (garbage-in, garbage-out).

At this point, the question is: Are there some special conditions under which residuals can be treated as inputs? The answer to this question has been provided in Ebert and Welsch (2007), from which most of the following developments are derived.

Let us explicitly split the input into material inputs x_M and service inputs x_S. Ebert and Welsch (2007) have considered the following representation:

$$y = F(x_M, x_S), x_M = y + z \tag{20}$$

where F satisfies the usual productivity axioms (Chambers 1988) accounting of course for the limitations introduced by the mass balance equation. Considering the mass balance equation, the production system in (20) can be transformed into:

$$y = F(y + z, x_S) \Leftrightarrow y = G(z, x_S), \frac{\partial G}{\partial x_S} \geq 0, \frac{\partial G}{\partial z} \geq 0 \tag{21}$$

The developments in (21) are possible by considering y as an implicit function G of z and x_S. In the last equation in (21), residuals z play the role of input variable (see Ebert and Welsch 2007 for more discussion of the properties of the function G).

The difference between the representations in (21) and (8) is that in the former material inputs are replaced by the residual levels, while in the latter both material inputs and residuals appear in the transformation function. It seems that in model (8), there is a double accounting of material inputs given that residuals are intrinsically related to those inputs.[5]

A particular case of the model in (21) can be obtained when there are no service inputs in the technology described by the function G. Then the good output is a function of only the residuals: $y = G(z)$. This case is

[5]This situation particularly makes sense since the relation describing how bad outputs are generated is missing.

reminiscent of the frontier eco-efficiency (FEE) discussed in Kortelainen and Kuosmanen (2004) and Kuosmanen and Kortelainen (2005), where residuals (environmental damages) are used as input to explain the value added.[6] The traditional definition of eco-efficiency is the ratio of value added to environmental damages. It links the ecological and economic outcomes, abstracting the conventional relations between inputs and outputs. In other words, the economic outcome is maximized given the environmental impacts, or, conversely, the environmental by-production is minimized given the economic outcome (profit, added value).[7]

Finally, though Ebert and Welsch (2007) have proved that residuals can be treated as inputs, their model representation still does not explicitly show how residuals are actually generated. Therefore, substituting for x_M may also be a misuse of the materials balance. Besides, describing a production process will show residuals as outputs of the transformation and not inputs; and for an economic actor, if residuals are considered as input, then their optimal level can be zero given other input quantities. Hereby, we can conclude that considering residuals as inputs actually goes against the materials balance principle.

Residuals as Outputs

In this part, we discuss residuals as outputs under the classic strong disposability assumption and also under the WDA (Shephard and Färe 1974). Let us redefine the graph technology and the input and output requirement sets as follows:

$$\begin{aligned} \Psi &= \left\{(x,y,z) \in \mathbb{R}_+^{K+Q+R} | x \text{ can produce } (y,z)\right\} \\ L : \mathbb{R}_+^{Q+R} \rightarrow L(y,z) &= \left\{x \in \mathbb{R}_+^K | (x,y,z) \in \Psi\right\} \\ P : \mathbb{R}_+^K \rightarrow P(x) &= \left\{(y,z) \in \mathbb{R}_+^{Q+R} | (x,y,z) \in \Psi\right\} \end{aligned} \tag{22}$$

Under the strong disposability assumption, the residuals are treated as another good output. The single structure of production technology has the maximum degree of assortment, defined by Frisch (1965) as the ability

[6]Lauwers (2009) has considered FEE models as a special case of pollution-generating technology modelling. Moreover, he has referred to the model presented in Eq. (17) as environmentally adjusted production efficiency models.

[7]The genesis of the FEE model can be found in Tyteca (1996, 1997).

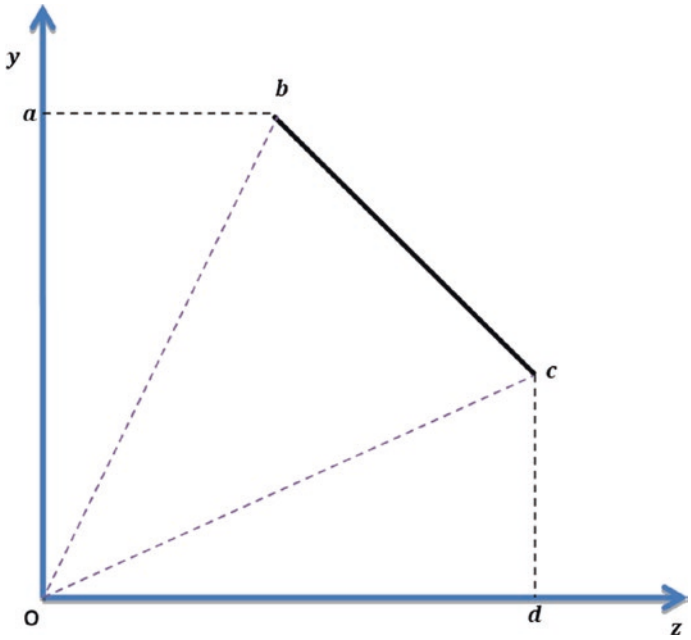

Fig. 2 Outputs isoquant representation under strong and weak disposability assumptions

to divert inputs to the production of any output without the generation of extra costs. Therefore, given this maximum flexibility the levels of residuals can be zero, which actually contradicts the mass balance equation. In Fig. 2, we have plotted the output isoquant and under the free disposability assumption the isoquant is represented by (*OabcdO*), which shows the possibility of zero residuals given some fixed levels of input.

A Shephard output distance function representing the technology will imply under this case an equiproportionate increase in both residuals and desirable outputs, which controverts the idea that residuals—especially bad environmental outputs—generate negative externalities and need to be reduced. The Shephard output distance function is defined as:

$$D_o(x, y, z) = \inf \left\{ \lambda : \left(\frac{y}{\lambda}, \frac{z}{\lambda} \right) \in P(x) \right\} \tag{23}$$

Using the mass balance equation in (1) and the Shephard output distance function, we have:

$$x_M = \frac{y}{\lambda} + \frac{z}{\lambda} \tag{24}$$

Again, the only solution to problem (24) is when $\lambda = 1$ and hence no inefficiency can be measured. As previously, the material input levels are held fixed and the strong disposability assumption of these inputs is not accounted for.

In terms of trade-offs, if we consider the transformation in (25):

$$F(x, y, z) = 0, F_y = \frac{\partial F}{\partial y} \geq 0, F_x = \frac{\partial F}{\partial x} \leq 0, F_z = \frac{\partial F}{\partial z} \geq 0 \tag{25}$$

Using implicit function theorem, it is easy to show that for a fixed level of inputs $\frac{dy}{dz} \leq 0$, which is coherent with the mass condition as demonstrated in (12). When desirable outputs are fixed, we have the following relation: $\frac{dz}{dx} \geq 0$, which is true for all the inputs (materials and services) yet we know that dematerialization through substitution of service inputs to material ones can help mitigate the level of residuals. In the absence of service inputs, the relation $\frac{dz}{dx} \geq 0$ is consistent with the mass balance equation if we totally differentiate the equation in (1) under a fixed level of the intended output. However, as we show using the Shephard output distance function, treating residuals similarly as another good output violates the idea that residuals create social costs that need to be alleviated.

To overcome the previous issues, Färe et al. (1986) have proposed the WDA to deal with residuals in addition to the null-jointness property in (15). In this peculiar situation, WDA means that reducing undesirable outputs requires decreasing the desirable output quantities by the same factor (Färe et al. 1996). Formally, the WDA can be summarized as:

$$(y, z) \in P(x), 0 \leq \gamma \leq 1 \Rightarrow (\gamma y, \gamma z) \in P(x) \tag{26}$$

Graphically, the WDA disposability of both residuals and intended-outputs restrains the output requirement set in Fig. 2 to (*ObcO*). However, the strong disposability of intended outputs is generally maintained. Then the output set is defined by (*ObcdO*). The WDA property presented in (26) is based on fixed levels of input and using the mass balance equation in (1) will imply $\gamma = 1$. In terms of distance function, Färe et al. (1989) estimated the hyperbolic efficiency which allows a simultaneous equiproportionate expansion of the good outputs and a contraction of the bad outputs and inputs by the same radial factor. Yet Coelli et al. (2007) prove the inconsistency of this measure regarding the mass balance equation (the efficiency equals one as a solution).

The hyperbolic distance function can be represented by:

$$D_h(x, y, z) = \inf \left\{ \lambda : \left(\lambda x, \frac{y}{\lambda}, \lambda z \right) \in \Psi \right\} \tag{27}$$

Using the mass balance equation and the hyperbolic distance function yields:

$$\begin{aligned} &\lambda x_M = \frac{y}{\lambda} + \lambda z \Leftrightarrow \\ &\lambda^2 (x_M - z) = y \\ &\Rightarrow \lambda = 1 \end{aligned} \tag{28}$$

Again, no inefficiency can be observed. Clearly, the hyperbolic efficiency estimation in (27) goes against the materials balance. Let us go one step further and introduce explicitly the WDA in the formulation in (28). We have:

$$\begin{aligned} &\lambda x_M = \frac{\gamma y}{\lambda} + \lambda \gamma z \Leftrightarrow \\ &x_M = \frac{\gamma y}{\lambda^2} + \gamma z \\ &\Rightarrow \lambda = 1, \gamma = 1 \end{aligned} \tag{29}$$

The previous demonstrations prove that the WDA may not be consistent with materials balance unless very restrictive condition where all observations are efficient and the abatement factor (scaling factor or disposability parameter) γ equals unity. The production technology under the WDA has also been described using a directional distance function (Chung et al. 1997; Färe and Grosskopf 2004b) and the inconsistency with the materials balance is discussed in Hoang and Coelli (2011). Alternative distance functions to (27) can be proposed, for example:

$$D_a(x, y, z) = \inf \left\{ \lambda : \left(x, \frac{y}{\lambda}, \lambda z \right) \in \Psi \right\} \tag{30}$$

Introducing this distance function in the mass balance equation yields:

$$\begin{aligned} &x_M = \frac{y}{\lambda} + \lambda b = y + b \\ &\Rightarrow \lambda = 1 \end{aligned} \tag{31}$$

If we consider that in (30), material inputs are actually strongly disposable, we have:

$$\hat{x}_M = \frac{y}{\lambda} + \lambda z \geq x_M = y + z \tag{32}$$

This last demonstration is consistent with the materials balance with the introduction of inefficiency in the use of material inputs. However, this is

not a strategy for a rational producer to become inefficient in order to comply with the materials balance principle.

In terms of trade-offs, the isoquant in Fig. 2 shows that the relation between residuals and intended outputs can alternatively be positive (along *Ob*), negative (along *bc*) or zero (along *cd*). See Dakpo et al. (2017b) for an extensive discussion of the limits of the WDA. Besides, most of the limits discussed for the case where residuals are treated as inputs or as other good outputs are also valid here. Along the *Ob* on the isoquant, residuals are like inputs, and along *bc* they are like another good output.[8] Despite those limitations, the WDA is the most widely used approach for benchmarking under a pollution-generating technology (Färe et al. 1996, 2001a, 2005, 2007; Weber and Domazlicky 2001; Arocena and Waddams Price 2002; Boyd et al. 2002; Lee et al. 2002; Domazlicky and Weber 2004; Zaim 2004; Picazo-Tadeo et al. 2005; Yörük and Zaim 2005; Kumar 2006; Marklund and Samakovlis 2007; Watanabe and Tanaka 2007; Lozano and Gutiérrez 2008; Picazo-Tadeo and Prior 2009; Kumar Mandal and Madheswaran 2010; Sahoo et al. 2011; Lee et al. 2016; Kao and Hwang 2017; Shen et al. 2017), even though Leleu (2013) has shown that many studies have inappropriately specified this property (incorrect linearization in the case of DEA). Some extensions of the WDA have even been discussed in the literature (Kuosmanen 2005; Zhou et al. 2008a; Kuosmanen and Podinovski 2009; Yang and Pollitt 2010; Kuosmanen and Kazemi Matin 2011; Podinovski and Kuosmanen 2011; Valadkhani et al. 2016; Chen et al. 2017; Pham and Zelenyuk 2018; Roshdi et al. 2018).

In the DEA literature, to address the issues associated with negative/positive shadow prices of residuals due to the WDA, many studies have used non-radial or slack-based measures (SBM) that account for all the sources of inefficiency and allow to compute the inefficiency separately for each input and output. They are termed non-radial measures and are considered 'complete' since they take into consideration all types of inefficiencies (Cooper et al. 1999a, b). For instance, one can find non-radial directional distance functions (Chang and Hu 2010; Zhou et al. 2012; Zhang et al. 2013); a weighted Russell directional distance model (WRDDM) in

[8]The fundamental problem with Fig. 2 is that when the bad z is treated as a normal good, then the obvious result is that a zero level of z can be realized. However, the crucial point is that the relation showing how z is generated is missing. The apparent trade-off between y and z is therefore an illusion and goes against the materials balance. The bad output is generated using input x_M that is constant along the output transformation curve implying that y cannot be increased and z decreased. This fact is independent of whether disposability is strong or weak, it simply follows from how the bad output is generated (Førsund 2018).

Chen et al. (2011) and Barros et al. (2012); the additive efficiency index (AEI) in Chen and Delmas (2012); range adjusted measures (RAM) in Sueyoshi and Goto (2010, 2011b) and Sueyoshi et al. (2010); median adjusted measures (MAM) in Chen (2013); and SBM in Tone (2004), Lozano and Gutiérrez (2011), and Song et al. (2014). Despite the interesting features of these models, at some point they are all equivalent to models that treat residuals as inputs and therefore they suffer from the same trade-off limits. Finally, in this framework of nonparametric DEA, Wang et al. (2012) have suggested to consider undesirable outputs as fixed outputs in the modelling.

Iso-Environmental Lines

Coelli et al. (2005) treated residuals not as inputs or outputs, but more as an 'outcome'. We choose this terminology here to specify that residuals are the results of a behaviour/attitude like in the case of profit, revenue maximization or cost minimization. 'Outcomes' here refer to an impact. The idea has been further elaborated in Coelli et al. (2007) and is based on earlier unpublished empirical work by Lauwers et al. (1999) on nutrient balance in pig production. The approach does not require the introduction of a residual variable contrary to all previous models (whether as an additional input or as an additional output). On the contrary, it relies on a mass balance equation as defined in (1) and (2). Since the objective is to minimize the 'surplus' $z = a'x - r'y$, the problem can be resolved analogously to a cost minimization programme. Many papers/empirical applications are based on the null vector for r, assuming that the desirable outputs do not contain any materials (Rødseth 2013; Rødseth and Romstad 2013; Guesmi and Serra 2015; Wang et al. 2018a). However, this may be misleading and it is better to assume that the levels of desirable outputs are given.

The problem that needs to be solved is then presented as:

$$N(y,a) = \min_{x}\{a'x | (x,y) \in \Psi\} \tag{33}$$

where $\Psi = \{(x,y) \in \mathbb{R}_{+}^{K+Q} | x \text{ can produce } y\}$. The boundary of $N(y,a)$ is equivalent to what is called an 'iso-environmental cost line' (in allusion to the commonly known iso-cost line). Based on this concept, environmental efficiency can be decomposed into technical efficiency (classic measure) and allocative efficiency (inefficiency due to a mix in material inputs). As in the case of a cost function, N is function of the desirable outputs and the input flow coefficient a. $N(y,a)$ thereby satisfies the following properties:

- Non-decreasing in a and y. if $a' > a$ then $N(y, a') > N(y, a)$, if $y' > y$ then $N(y', a) > N(y, a)$.
- Homogeneous of degree 1 in a. $N(y, ta) = tN(y, a)$ for $t > 0$.
- Concavity in a. $N(y, ta + (1 - t)a') \geq tN(y, a) + (1 - t)N(y, a')$.
- Continuity in a.

The first property guarantees a positive correlation between residuals and good outputs. If price information is available, one can also identify the cost $(w'x_e)$ of the input bundle x_e that minimizes the residual surplus $N(y, a)$ and, in addition, determine the residual level associated with a cost minimization programme $(a'x_{\min})$. The difference between the two costs values $(w'x_e - w'x_{\min})$ is the 'shadow cost' of pollution reduction.[9] An advantage of this framework of environmental performance assessment is that it can be helpful to determine a tax level $(\alpha > 0)$ aiming at residual reduction [in fact the two iso-lines (cost and environmental) coincide when $w = \alpha a$]. Another interesting aspect of this approach is that, depending on the region of the isoquant, some improvements towards the environmental efficiency point can result in a decrease of the costs. For instance in Fig. 3, we have represented the two iso-lines considering two material inputs x_1 and x_2. The environmental efficiency for the observation A is OB/OA and the corresponding costs are lower in point B. Of course this is not always the case since it depends on 'whether the movement is towards or away from the cost minimizing point' (Coelli et al. 2005).

Within this framework of environmental economics, the materials balance principle offers four different ways of reducing residuals: (i) improving technical efficiency (which is cost reducing), (ii) increasing environmental allocative efficiency (the effect of which depends on the direction),[10] (iii) using extra inputs for pollution abatement (which is costly) and (iv) output reduction (which implies profit losses). Applications of the materials balance principle can be found in Welch and Barnum (2009), Van Meensel et al. (2010), Hoang and Alauddin (2011), Hoang and Nguyen (2013), Kuosmanen and Kuosmanen (2013), Kuosmanen (2014), Aldanondo-Ochoa et al. (2017), and Wang et al. (2018b). A SFA of the materials balance principle has also been discussed by Hampf (2015).

[9] This cost can be lowered using abatement options. The pollution excess then equals $a'x_{\min} - a'x_e$.

[10] This idea suggests substituting high emission factor inputs with ones with low emission factor or substituting low recuperation factor outputs with ones with high recuperation factor.

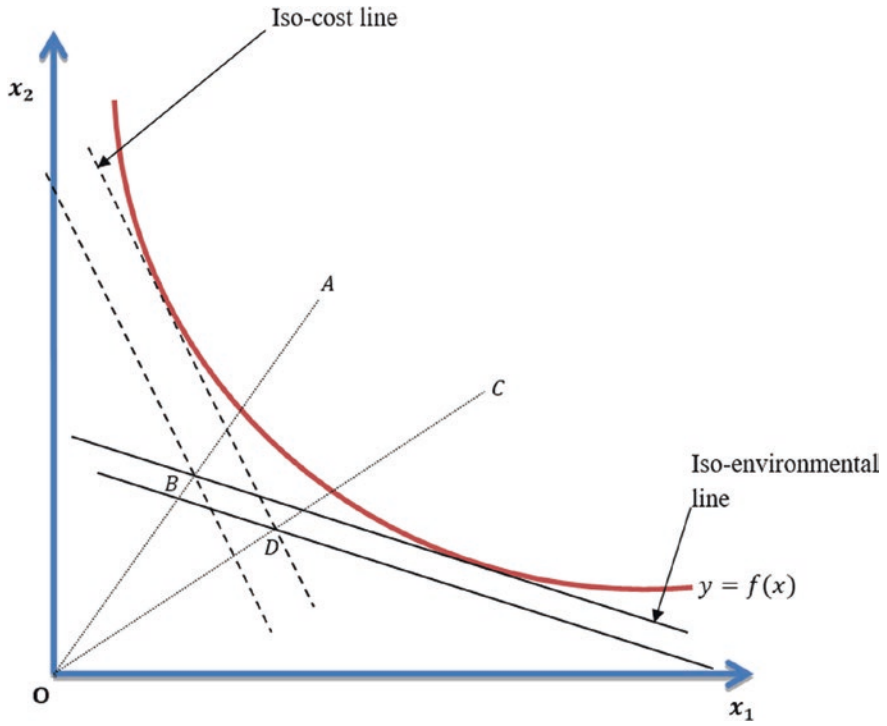

Fig. 3 Iso-cost and iso-environmental line representations (*Source* Adapted from Coelli et al. 2007, p. 7)

In the same line of Ebert and Welsch (2007) and Hoang and Rao (2010) formulated two main drawbacks of the traditional materials balance principle: (i) 'ambiguous treatment of immaterial inputs' and (ii) 'lack of universally accepted weights for various materials'. According to these authors, the minimization of problem (33) ignores the existence of non-material inputs that do not fall in the range of residuals-generating inputs and therefore have zero material contents. Besides, the problem in (33) can only deal with one type of residual (or impact) unless some weights can be provided to aggregate different kind of impacts (Dakpo et al. 2017b). Modelling residuals as outcomes like costs, revenues or profits, implies that these residuals are not 'immediately observable' or measurable. They need to be estimated to some extent. While this is true for many residuals, such as non-point source pollution, other residuals can be directly measured at the firm level, for example, smoke emission at a plant level, toxic releases in water, etc. Even if residuals are considered outcomes rather than outputs, we still do not know how they are generated within this framework.

The Weak G-Disposability

Acknowledging the importance of the materials balance principle, Hampf and Rødseth (2015) and Rødseth (2015) have recently introduced the weak G-disposability as a restriction of the G-disposability discussed by Chung et al. (1997). If the technology is defined as in (22), the G-disposability implies that:

$$(x, y, z) \in \Psi \Rightarrow (x + g_x, y - g_y, z + g_z) \in \Psi \tag{34}$$

where g_x, g_y, g_z are directional vectors. The weak version of the G-disposability constrains the directional vectors using a summing-up condition:

$$a' g_x + r' g_y - c' g_z = 0 \tag{35}$$

Using the mass balance equation in formula (2), we have:

$$\begin{aligned} a'(x + g_x) = r'(y - g_y) + c'(z + g_z) \Rightarrow \\ a' g_x + r' g_y - c' g_z = 0 \end{aligned} \tag{36}$$

The summing-up condition is therefore a restriction to make the technology representation to comply with the materials balance principle.

Following Hampf and Rødseth (2015) and Rødseth (2015), the technology Ψ should verify the following postulates:

MB1: No free lunch
MB2: Non-emptiness
MB3: Closedness
MB4: The output set $P(x)$ is bounded
MB5: Convexity
MB6: Output essentiality for the unintended outputs: $(x, y, z) \in \Psi$, $z = 0 \Rightarrow x_M = 0$, where x_M represent the residuals-generating inputs
MB7: Input essentiality for the unintended outputs: $(x, y, z) \in \Psi$, $x_M = 0 \Rightarrow b = 0$
MB8: Weak G-disposability of inputs and outputs: $(x, y, z) \in \Psi$, $a' g_x + r' g_y - c' g_z = 0 \Rightarrow (x + g_x, y - g_y, z + g_z) \in \Psi$ where $g_.$ are direction vectors
MB9: Returns to scale assumptions.

Under the postulates MB6 and MB7, the second law of thermodynamics is verified. As opposed to the weak disposability model (under the WDA) inputs are no longer freely disposable. As a matter of fact, under the WDA, the free disposability of inputs stipulates that for a given input bundle and a produced output set (including good and residual outputs), it is possible for a higher input bundle to produce the same amount of the output set. However, this is technically infeasible under the conditions of the materials balance principle (especially for the residuals).

Figure 4 shows two DMUs (A, B) that use the same non-zero amount of inputs to produce one good output and one residual. For this figure, we have assumed zero recuperation factors for the good outputs. Under the WDA, the residual level that can be generated is zero (meaning that residuals are not essential to the consumption of material inputs) because of the null-jointness assumption, while under the weak G-disposability this minimum can be found in point D. More explicitly, under the WDA, the output technology set spans over $(OABCO)$, but under the materials balance principle, the output set is $(DABCD)$, which is narrower than the one under the WDA. The vertical lines capture the inelasticity between good and undesirable outputs since we have posited no recuperation factors. The segment $[AB]$ reflects the convexity assumption. It is worth noting that in Fig. 4 the good output intensity per bad output is higher in point A than in point B. However, under the materials balance principle and variable returns to scale, the technology Ψ is represented by $(DABCD)$ while under constant returns to scale (CRS) the technology is displayed by all points on the east side of (DAE).

Only a few applications of the weak G-disposability can be found in the literature (Rødseth 2016; Cui and Li 2017b; Hampf and Rødseth 2017; Hampf 2018a, b; Wang et al. 2018a).

Again, the technology is restricted to satisfy the mass balance equation, and there is no description of how residuals are generated. Rigorously speaking, at the frontier the residual is considered as another input (see segment $[AB]$ in Fig. 4) under this single structural representation of the technology. This means that a functional trade-off between the good and the bad output is assumed and as we previously argued this crucial feature is incompatible with the materials balance principle. Methodologically, Hampf and Rødseth (2015) have shown that under some conditions, the weak G-disposability is equivalent to the WDA and therefore it suffers from the same limitations. Finally, the summing-up condition imposes some specific constraints on the disposability of the different variables. A producer may not be aware of those, and therefore, his decisions will not be made simply to comply with

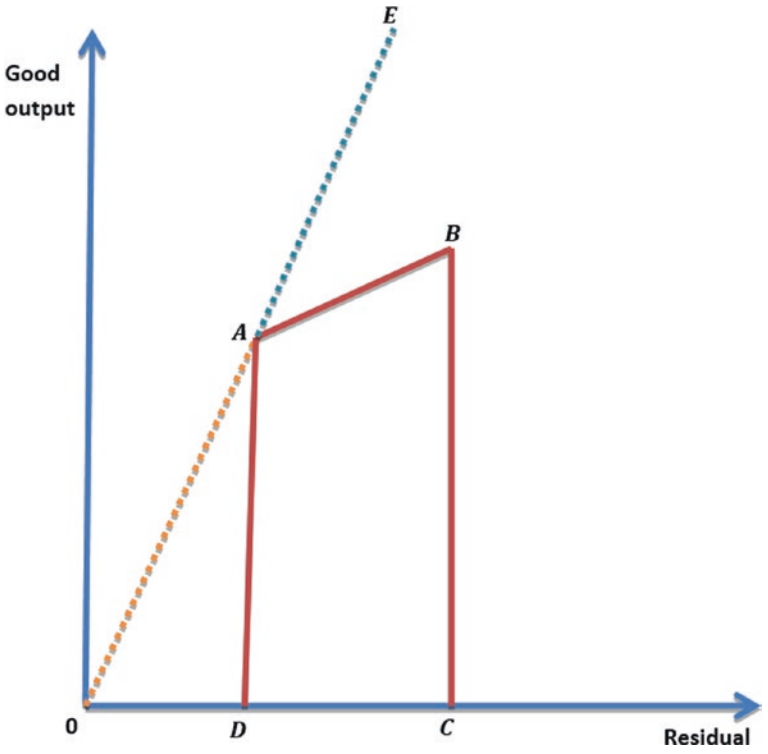

Fig. 4 Weak G-disposability representation (*Source* Adapted from Hampf and Rødseth 2015)

the materials balance principle. Therefore, it does not really help to have the materials balance constraint.

Data Transformation or Indirect Approaches

Among the single structure representations of production technologies including residuals, we have the data transformation models. The approaches proposed in this category are based on applying an asymmetric transformation function to the residuals and then considering a classic model with multiple outputs. A review of several transformation functions that can be used to deal with residuals can be found in Scheel (2001). Seiford and Zhu (2002) suggested for example a linear monotone decreasing function (additive inverse):

$$\bar{z}_r = -z_r + v \geq 0 \tag{37}$$

where v is a proper[11] translation vector.[12] The model can only be estimated under variable returns to scale.

This approach has recently been challenged by Färe and Grosskopf (2004a) through a comparison with the model under the WDA.[13] Their results showed the inconsistency of the data transformation approach. This is quite understandable since the transformation distorts the real production process. Moreover, the model implies that residuals are freely disposable and can be reduced without any cost, which is not realistic (Du et al. 2013). While many studies make assumptions based on the linear transformation function, others directly consider residuals as a negative output (Jahanshahloo et al. 2004; Zhu and Cook 2007; Fleishman et al. 2009; Liang et al. 2009). On the other hand, Lovell et al. (1995) and Scheel (2001) proposed to use the inverse function (multiplicative inverse):

$$\bar{z}_r = \frac{1}{z_r} \tag{38}$$

However, this model alters the production function, creating missing values (for zero values, the inverse does not exist), and thus may lead to questionable results.

Other applications and illustrations of the data transformation approach can be found in Vencheh et al. (2005), Hua et al. (2007), and Pérez et al. (2017).

In conclusion, single structure systems—even though they can be made compatible with the materials balance principle—are very limited in describing production technologies, especially in terms of the way residuals are generated.

3.2 Multi-equation Pollution-Generating Systems

Some Concepts

The concept of multi-equation production was introduced in the economic literature by Frisch (1965) to describe complex transformation systems involving several technically connected products for which inputs are jointly

[11] v is sufficiently large so that the new variable is positive.

[12] For the case of the banking industry, Berg et al. (1992, p. 219) have used the additive inverse where $v = 0$ to introduce loans losses as negative outputs.

[13] See also Liu and Sharp (1999) for further discussion on issues related to data transformation.

or alternatively used. According to Frisch, a production system can be represented by multiple functional relations, each one describing the relations between inputs and outputs and among inputs or outputs. Frisch (1965) has defined some concepts that help understand this type of modelling. Considering a system that produces Q different outputs related by μ (independent) functional relations between outputs and or inputs, the degree of assortment (freedom) of the system equals $\alpha = Q - \mu$. When $\mu = 1$, we have a standard representation of the production technology with a single production relation and maximum degree of assortment. In this case, given the maximal flexibility of this system, inputs can be directed towards the production of any outputs without generating additional costs. It is possible to have some relationships between inputs, independently of outputs. These relations have been coined pure factor bands and the degree of assortment can be negative in their presence. When the degree of assortment equals zero, we have factorially determined multi-equation production. In factorially determined technologies, given the level of inputs, all the outputs are determined. Irrespective of Q, μ, α, relations involving only product quantities (pure product bands) can be present in the system. The number of these relations (κ) represents the degree of coupling (factor free) of a multi-equation production. κ is simply the number of output relations that can be deduced from μ independently from the production factors. Both the degree of assortment and coupling determine the level of flexibility present in a system, i.e. the flexibility in the determination of output mix. For instance, a higher assortment corresponds to a higher flexibility and a degree of coupling equal to zero also implies flexibility.

The By-Production Approach

Initial adaptation of the multi-equation production system to the case of residual generation (environmental externalities) can be found in Førsund (1972, 1973) with illustrations of general equilibrium models. However, it took more than three decades for the ideas put forward in Frisch (1965) to be extended to the performance benchmarking of firms. In this framework, Førsund (1998) is to our knowledge the first to use the Frisch concept to criticize the treatment of residuals as inputs or as outputs under the WDA. Later, Murty and Russell (2002) and Murty et al. (2012) have introduced the by-production approach as a formal way to treat residuals under the multi-equation production system, even though these authors have not explicitly referred to the concepts defined in the previous sub-section.

In this line, Murty (2010b) has defined five attributes that characterize pollution-generating technologies:

- The use of material inputs will necessarily result in the generation of residuals (wastes, pollution). Murty (2010b) has referred to this as nature's emission generating mechanism[14] that is triggered when one uses emission-causing inputs. We believe that this attribute is very close to the material inputs essentiality previously defined.
- Since the idea of the by-production is to explicitly model how residuals are generated, a distinction must be made between inputs destined to the production of intended outputs and those that generate residuals. Those inputs are non-rival or joint because their use in the production of intended outputs does not prevent them from generating residuals. This attribute is the essence of Frisch (1965) ideas where each output is described by its own production relation. Thereby, the production system is described by two sub-technologies, one producing intended outputs and the other generating residuals.
- Residuals are not freely disposable but costly to dispose (Murty 2010a). The costly disposability implies that for each level of emission-causing input, there is a minimal level of residuals that can be generated and more than this minimal level is possible in the presence of inefficiency. Under this property, the function describing the shape of the frontier of residual generation is convex. Costly disposability is simply the polar opposite[15] of the standard free disposability assumption. Moreover, strong disposability is maintained for service inputs and intended outputs, while pollution-generating goods violate the free disposability assumption. At this point, it is worth mentioning that in some cases some intended outputs can generate negative externalities; however, we do not deal with those cases in this chapter. Recently, Murty (2015) has extended the free and costly disposability to conditional versions. Conditional free disposability refers to the changes in the minimal amount of residuals given that higher levels of material inputs are feasible under the intended output production sub-technology. The conditional costly disposability assumption implies the opposite, i.e. with lower levels of material inputs feasible

[14]Even if the idea around this concept is clear, it is worth mentioning that it is very strange to quote 'nature's emission generating mechanism'.

[15]We believe that the expression 'polar opposite' used by Murty et al. (2012) simply refers to a systematic reverse in the inequalities. This means that free disposability and costly disposability are complete/exact opposites of one another.

under the residuals-generating sub-technology, the maximum amount of intended outputs has to change consequently.

- A positive correlation between residuals and intended outputs is a direct consequence of all the previous attributes. Any increase in material inputs will result in an increase in both intended outputs and residuals.
- The fifth attribute is related to the treatment of abatement or cleaning activities. According to Murty (2010a) and several other papers (Kumar and Managi 2011; Färe et al. 2012; Yu-Ying Lin et al. 2013), resources can be diverted from the production of intended outputs to the generation of abatement outputs useful in mitigating residuals. The cost of this resource diversion is the production of less intended outputs. Under this approach, both intended and abatement outputs are treated under the same activity (same technological process). In other words, intended outputs and abatement outputs cannot be identified separately.

The specific case of by-production discussed in Murty et al. (2012) [MRL hereafter] described the global technology as the intersection of two sub-technologies: one dedicated to the production of goods and the other one to the generation of bads. The global technology can be specified as follows:

$$\Psi = \Psi_1 \cap \Psi_2 = \left\{ \begin{array}{c} (x_M, x_S, y, z) \in \mathbb{R}_+^{\mathrm{K}_M+\mathrm{K}_S+\mathrm{Q}+\mathrm{R}} \,|\, f(x_\mathrm{M}, x_\mathrm{S}, y) \leq 0, \\ z \geq g(x_\mathrm{M}) \end{array} \right\} \tag{39}$$

where

$$\Psi_1 = \left\{ (x_M, x_S, y, z) \in \mathbb{R}_+^{K_M+K_S+Q+R} | f(x_M, x_S, y) \leq 0 \right\} \tag{40}$$

$$\Psi_2 = \left\{ (x_M, x_S, y, z) \in \mathbb{R}_+^{K_M+K_S+Q+R} | z \geq g(x_M) \right\} \tag{41}$$

and f and g are both continuously differentiable functions. At this point, a relevant fact that has not been clearly mentioned by Murty et al. (2012) is whether the global technology defined as the intersection of two sub-technologies can be empty. The answer to this question is directly linked to the mass balance equation in (1) or (2), where for each level of material input there will always be a (minimal) corresponding value of pollution generated through the production process. In other words, as long as the materials balance principle is satisfied, the global technology can never be empty.

The intended-outputs sub-technology satisfies the standard disposability assumptions:

$$(x_M, x_S, y, z) \in \Psi_1, \tilde{x}_M \geq x_M, \tilde{x}_S \geq x_S, \tilde{y} \leq y \Rightarrow (\tilde{x}_M, \tilde{x}_S, \tilde{y}, z) \in \Psi_1 \tag{42}$$

For the transformation function in (40), the disposability assumptions in (42) imply:

$$\frac{\partial f}{\partial x_{\mathrm{M}}} \leq 0, \frac{\partial f}{\partial x_{\mathrm{S}}} \leq 0, \frac{\partial f}{\partial y} \leq 0 \tag{43}$$

The costly disposability assumption with respect to the residuals can be expressed as follows:

$$(x_M, x_S, y, z) \in \Psi_2, \bar{z} \geq z, \bar{x}_M \leq x_M \Rightarrow (\bar{x}_M, x_S, y, \bar{z}) \in \Psi_2 \tag{44}$$

The costly disposability implies that it is possible to generate more residuals given the levels of material inputs x_M, i.e. that the set of technology Ψ_2 is bounded below (Fig. 5) (Murty 2010a). This implies that:

$$\frac{\partial g}{\partial x_M} \geq 0 \tag{45}$$

Using implicit function theorem, by definition we have $x_M = h(z)$ and then one can write for an efficient observation:

$$f(h(z), x_S, y) = 0 \tag{46}$$

The trade-off between residuals and intended outputs is then[16]:

$$\frac{\mathrm{d}y}{\mathrm{d}z} = -\frac{\partial f}{\partial x_M}\frac{\partial h}{\partial z} \geq 0 \tag{47}$$

Following MRL, through changes in material inputs, the trade-off between intended outputs and residuals is positive, which is consistent with the mass balance equation as shown in (13). The by-production case described here

[16]Using the same ideas of trade-off estimation, MRL also proved the issues related to the single structure representation of a pollution-generating technology.

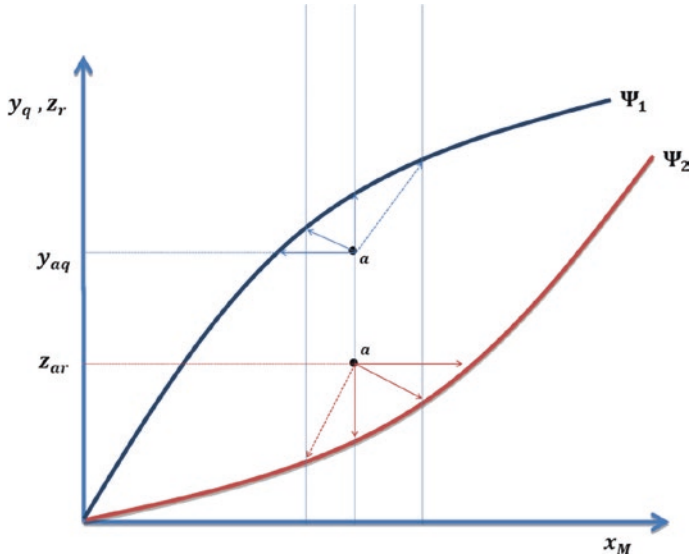

Fig. 5 The by-production technology representation (*Source* Adapted from Sueyoshi and Goto 2010)

is one of many possible cases. Murty (2015) and Murty and Russell (2016) demonstrated and argued that modelling residuals-generating technology has to be done case-by-case and that multiple models exist. Depending on the case under consideration, one may require multiple production relations to describe the overall technology (see for instance Levkoff 2013). However, in every case, one first describes the real-world production relations that correspond to the case at hand.

Early applications of MRL's by-production approach can be found in Chambers et al. (2014) and Serra et al. (2014). Lozano (2015) has also applied the by-production approach but described the technology as a parallel-processes network. Other applications can also be found in Guesmi and Serra (2015), Malikov et al. (2015), Kumbhakar and Tsionas (2016), Cui and Li (2017a), Dakpo et al. (2017a), Seufert et al. (2017), Zhao (2017), and Arjomandi et al. (2018).[17]

In the same line as Murty et al. (2012), Sueyoshi et al. (2010), and Sueyoshi and Goto (2010) also defined two efficiency models: an

[17]In the stochastic frontier framework, the idea of describing production systems using separable technologies for intended and unintended outputs has also been discussed in Fernández et al. (2002), Fernández et al. (2005), and Malikov et al. (2018).

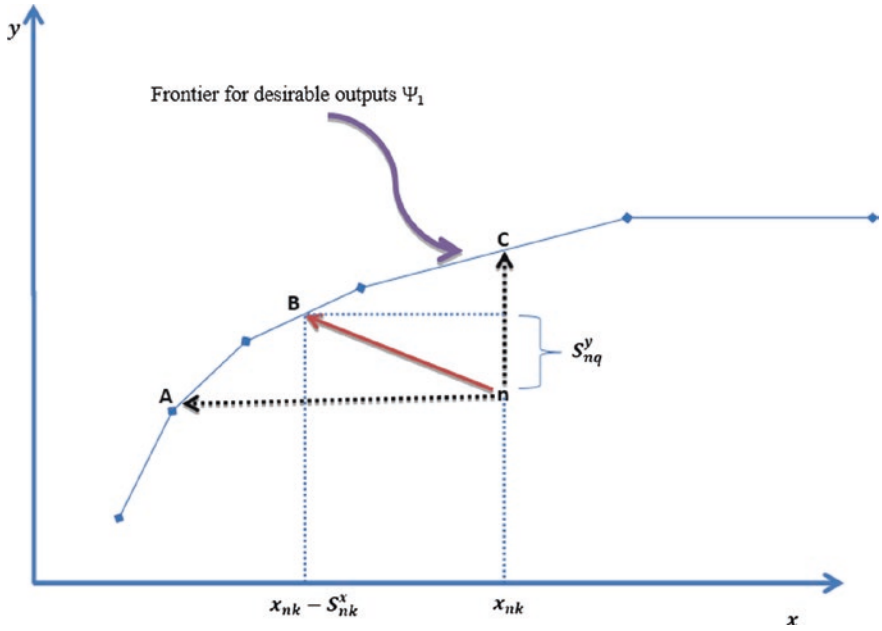

Fig. 6 Operational performance (*Source* Adapted from Sueyoshi and Goto 2010)

operational performance model and an environmental performance model. Their work can be related to the by-production approach as their operational performance concept is based on sub-technology Ψ_1 while the environmental performance concept is related to sub-technology Ψ_2. Sueyoshi et al. (2010) did not consider input separation (all the inputs cause residuals). The operational performance can be easily estimated as shown in Fig. 6, where an inefficient DMU_n can be projected towards the frontier for instance on point B (in Fig. 6, S represents the slacks).

For the environmental performance, Sueyoshi et al. (2010) and Sueyoshi and Goto (2010) have defined two disposability concepts aiming at analysing the 'adaptive behaviors' of DMUs in the presence of environmental regulations. The first concept is natural disposability (negative adaptation), where the manager chooses to reduce the consumption of inputs as the strategy for decreasing pollution (see Fig. 7 where the inefficient DMU_n is projected towards point J). The second concept is managerial disposability or positive adaptation (Sueyoshi and Goto 2012a, b, d), where managerial efforts, such as the adoption of cleaner technologies or the substitution of clean inputs for polluting ones, enable an increase in the consumption of inputs and simultaneously a reduction in pollution (see Fig. 8 where DMU_n

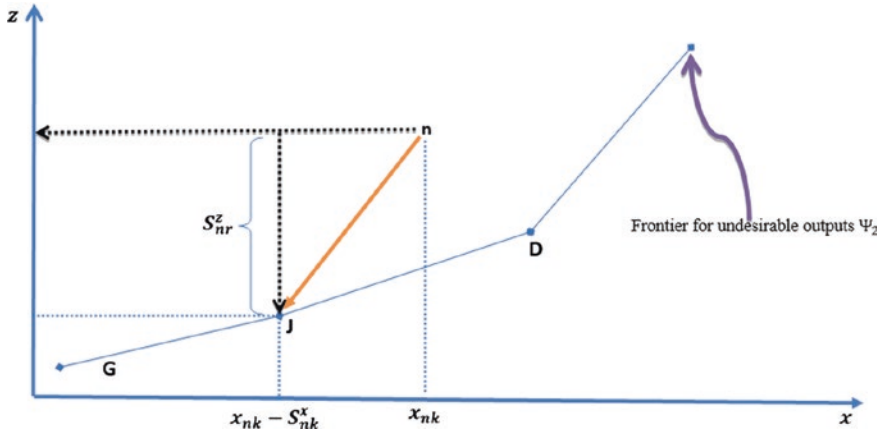

Fig. 7 Environmental performance under natural reduction of residuals

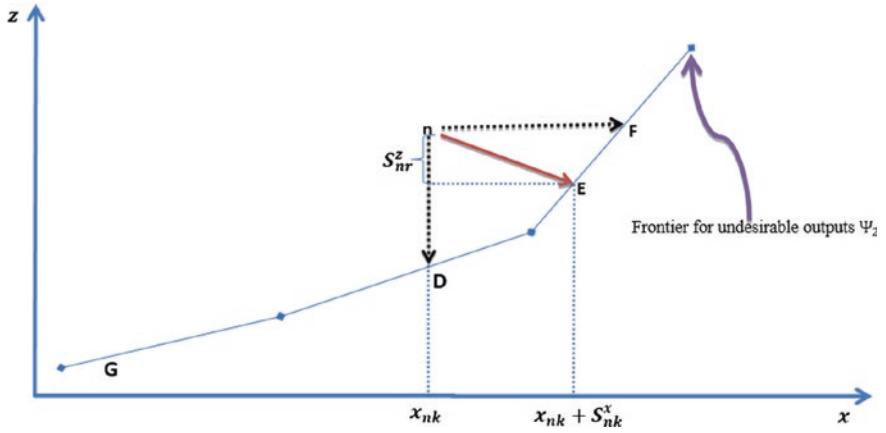

Fig. 8 Environmental performance under managerial effort

is projected on point *E*). This concept is grounded on the idea developed by Porter and van der Linde (1995) that regulation might offer innovation opportunities to secure the production of more good outputs and decrease the generation of bad outputs. However, from the economist's notion of economic efficiency-improvement, an inefficient production vector must imply increasing good outputs, reducing emission, while at the same time using fewer resources (inputs). More discussion on those disposability concepts can be found in Sueyoshi and Goto (2018a).

Those new disposability concepts have been largely used in many applications by the same authors and others (Sueyoshi and Wang 2014; Sueyoshi

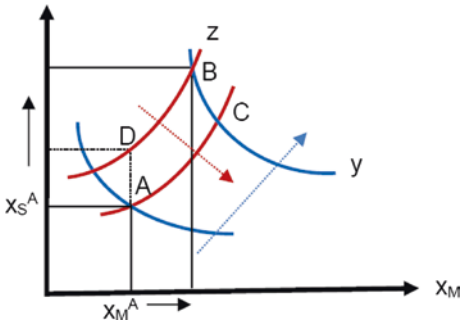

Fig. 9 Input isoquants representation (*Source* Taken from Førsund 2017)

and Yuan 2015, 2016) under the framework of DEA (Sueyoshi and Goto 2011a, b, c, 2015b, 2018b; Sueyoshi et al. 2013, 2017a, b, 2018; Cui and Li 2018; Sun et al. 2018). Fundamentally, the ideas that underlie the managerial and natural disposability are very similar to the ones previously proposed for the by-production. There are hardly any new concepts here, 'natural disposability' is reducing the use of inputs, and 'managerial disposability' is input substitution and the introduction of new waste-reducing technology. These two concepts are different and usually treated separately. However, their representations (unification) do not properly describe residual generation in terms of DEA representations. In addition, in this latter framework, they have sometimes introduced some non-linearity which may not identify efficient DMUs (Manello 2012).

On the other hand, Ray et al. (2017) have recently introduced a joint disposability between residuals and material inputs where the former cannot be reduced without reducing the latter. The joint disposability can be expressed as:

$$(x_M, x_S, y, b) \in \Psi_2, 0 \leq \alpha \leq 1 \Rightarrow (\alpha x_M, x_S, y, \alpha b) \in \Psi_2 \tag{48}$$

Nevertheless, the joint disposability only applies to the residuals sub-technology omitting the links with the intended-outputs sub-technology. This may result into the violation of the materials balance principle.

An important step has been brought forth by Førsund (2017), who argued that MRL representation of the technology is partly incomplete since dematerialization through substitution between material and service inputs is not covered (see Fig. 9 for the isoquant representation in the presence of substitution possibilities between the two type of inputs).

As such, he proposed to extend the residual-generating sub-technology as follows:

$$\Psi_2 = \{(x_M, x_S, y, z) \in \mathbb{R}_+^{K_M+K_S+Q+R} | z \geq g(x_M, x_S)\} \tag{49}$$

where the marginal productivity signs are:

$$\frac{\partial g}{\partial x_M} \geq 0, \frac{\partial g}{\partial x_S} \leq 0 \tag{50}$$

As argued in Førsund (2017), the multi-equation representation of a residual-generating production system is certainly one of the best strategies without an explicit inclusion of the mass balance equation. The model describes several processes that govern the production system without violating physical laws (or the materials balance principle). Moreover, the factorially determined multi-output model discussed in Førsund (2017) may crucially satisfy the materials balance where the single-equation model cannot.

4 Abatement Outputs (Technologies)

Throughout the previous discussion, we barely touched on abatement outputs that are possible with the adoption of end-of-pipe technologies. End-of-pipe technologies are pollution-control treatments aimed at improving the environmental performance of processes by partially removing already formed pollutants (Hellweg et al. 2005). As such solutions are physically separated from the production process, these type of technologies are generally located in the last stage of the system, before the disposal of the outputs, hence the name 'end-of-pipe technologies'. Since the installation of end-of-pipe technologies is a separate activity, they require the operationalization of their own inputs (materials and energy) (Zotter 2004). Examples of end-of-pipe technologies range from carbon dioxide removal or capture technologies (Olajire 2010; van Vuuren et al. 2018) to palm oil mill effluent management (Wu et al. 2010), incineration for waste disposal or wastewater treatment plants (Rennings et al. 2006). End-of-pipe technologies can be viewed as add-on packages ('react and treat') and should be clearly distinguished from clean production technologies (Glavič and Lukman 2007). The latter involve changes in the production process itself (process, product and organizational innovations) and therefore prevent the generation

of pollution during the production process (Sarkis and Cordeiro 2001; Rennings et al. 2006).

The main question remains: What are the incentives for the adoption of these abatement technologies? The literature shows that adoption is highly dependent on regulatory measures and the stringency of environmental policies (Frondel et al. 2007; Hammar and Löfgren 2010). While the adoption of cleaner integrated technologies is an important avenue for pollution mitigation, its relation with the modelling of pollution-generating technologies is beyond the scope of this chapter. Similarly, the short- and long-term effects of environmental policies (for instance the competitive advantage in relation to the porter hypothesis (Porter 1991; Porter and van der Linde 1995)) are also beyond the scope of this chapter. Our focus is rather on how to properly include abatement options in the modelling of pollution-generating technologies to consider the supplementary costs associated with these technologies.

In the presence of such outputs, the mass balance equation in (2) is written as:

$$a'x_M \equiv r'y + c'z^d + f'z^a \tag{51}$$

where z^a is the amount of pollution abated through end-of-pipe control technologies and z^d the residual pollution disposed in the environment. The primary amount of pollution is the addition of the two new subtypes ($z = c'z^d + f'z^a$). According to Rødseth (2014) and Hampf and Rødseth (2015), the WDA as proposed in Färe et al. (1989) is consistent with the materials balance principle only in the presence of end-of-pipe technologies. For the case of profit maximization, Rødseth and Romstad (2013) augment the classic production technology with the identity in (51) to account for abatement outputs. Later, Rødseth (2015, p. 3) argued that 'this is not a satisfactory result since the requirements on end-of-pipe abatement efforts are strong and, generally, physically unattainable'. Moreover, all these models are still framed in the single framework structure, with all of its limitations. Other papers deal with abatement activities using a network structure (Färe et al. 2013; Hampf 2013; Cui and Li 2016; Song et al. 2017; Bi et al. 2018). Nevertheless, the primary levels of pollution are treated in a single framework structure that again suffers from the aforementioned issues. In the case of by-production, Murty et al. (2012) have considered abatement output as another intended output which appears in both the intended and the unintended sub-technologies. However, in their formulation, abatement output through end-of-pipe technologies is still not properly modelled as a

separate activity. An adequate modelling of abatement output can be found in Førsund (2018).

5 Environmental Goods as Conventional Outputs in a Distance Function[18]

The previous sections have dealt with residuals as detrimental or unwanted outputs like pollution. However, many production processes also yield good environmental outputs, such as carbon sequestration in livestock farming. These good outputs are different from abatement outputs produced through end-of-pipe technologies. We believe that these outputs must also be accounted to gain a complete representation of an environmental technology that simultaneously generates bad and good environmental outputs.

5.1 Theoretical Background

Consider a firm that transforms a vector of $k = 1, \ldots, K$ inputs, $x \in \mathbb{R}_+^K$ to a vector of $q = 1, \ldots, Q$ outputs, $y \in \mathbb{R}_+^Q$. This transformation also yields a vector of $d = 1, \ldots, D$ environmental goods, $e \in \mathbb{R}_+^D$. In analogy to treating pollutants as inputs in the tradition of Baumol and Oates (1988), environmental goods are commonly assumed to have the same axiomatic properties as outputs. All feasible combinations of inputs, outputs and environmental goods (x, y, e) are characterized by the primitive technology set Ψ:

$$\Psi = \{(x, y, e) : x \text{ can produce } (y, e)\} \tag{52}$$

Ψ is assumed to be a closed and convex technology set with strongly disposable inputs, outputs and environmental goods. The corresponding output set is bounded. Most reviewed studies employ an output set, holding inputs constant. However, this keeps the relationship between inputs and environmental goods implicit. The primitive technology set encompasses the

[18]A similar, condensed argumentation with corresponding references can be found in the introduction of Ang, Mortimer, Areal and Tiffin (2018).

The reference to be added is:
Ang, F., S.M. Mortimer, F.J. Areal, and R. Tiffin, On the opportunity cost of crop diversification. *Journal of Agricultural Economics* 69 (3): 794–814 (2018).

And I suggest to replace the logical symbol for 'and' (an inverse V) by a comma. See formulas (8), (9), (21), (25), (29), MB6, MB7, MB8, (42)–(44), (48) and (50). Makes them simpler, like (26).

output set and makes this relationship explicit (Färe and Grosskopf 2005). Following Chambers et al. (1996, 1998), technology in (52) can be equally represented by the directional distance function:

$$\vec{D}_{\Psi}(x, y, e; g_x, g_y, g_e) = \max_{\beta}\{\beta : (x - \beta g_x, y + \beta g_y, e + \beta g_e) \in \Psi\} \tag{53}$$

where g_x, g_y and g_e are the directional vectors that specify the direction of, respectively, input contraction, output expansion and environmental good expansion towards the frontier. $\vec{D}_{\Psi}(.) \geq 0$ is differentiable and measures the distance to the frontier in the direction of (g_x, g_y, g_e).

The derivative of $\vec{D}_{\Psi}(.)$ with respect to outputs is:

$$\partial_y \vec{D}_{\Psi}(x, y, e; g_x, g_y, g_e) \leq 0 \tag{54}$$

The derivative of $\vec{D}_{\Psi}(.)$ with respect to inputs is:

$$\partial_x \vec{D}_{\Psi}(x, y, e; g_x, g_y, g_e) \leq 0 \tag{55}$$

The derivative of $\vec{D}_{\Psi}(.)$ with respect to environmental goods is:

$$\partial_e \vec{D}_{\Psi}(x, y, e; g_x, g_y, g_e) \leq 0 \tag{56}$$

Although environmental goods are non-marketed, we can assess the unknown shadow price u by exploiting the directional distance function's dual relationship to the profit function and by using the envelope theorem.

5.2 The Trade-Off Between Environmental Goods and Conventional Outputs

The trade-off between environmental goods and conventional outputs can be inferred using the envelope theorem (Chambers et al. 1996, 1998):

$$-\frac{\partial_e \vec{D}_{\Psi}(x, y, e; g_x, g_y, g_e)}{\partial_y \vec{D}_{\Psi}(x, y, e; g_x, g_y, g_e)} = -\frac{u}{p} \leq 0 \tag{57}$$

Equation (57) assumes that the shadow price u is positive and the relationship between marketable outputs and environmental goods is competitive for all levels of the environmental good. Figure 10 shows the production possibility frontier for one environmental good e_1 and one marketable

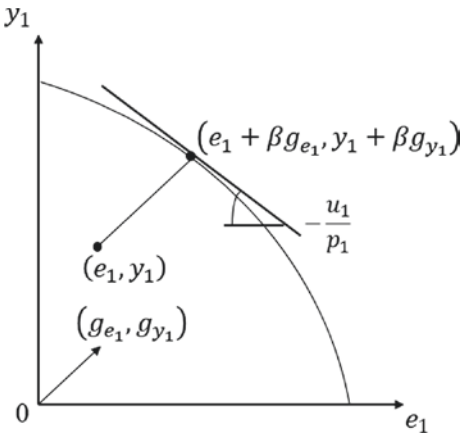

Fig. 10 Trade-off between one marketable output y_1 and one environmental good e_1, holding other outputs, environmental goods and inputs constant

output y_1, holding other environmental goods, as well as other outputs and inputs constant.

Färe et al. (2001b), Bellenger and Herlihy (2010), Ruijs et al. (2013), Sipilainen and Huhtala (2012), Bostian and Herlihy (2014), and Ruijs et al. (2015) use Eq. (58) to calculate the shadow price of environmental goods. To our surprise, only a handful of studies that use an augmented distance function discuss or check the assumption of a competitive relationship between marketable outputs and environmental goods in depth. Macpherson et al. (2010) conduct a correlation analysis and do not find a robust negative competitive relationship between the environmental goods and the marketable outputs. Sipilainen and Huhtala (2012) briefly mention that crop diversification has a private value, as it is a way to hedge against uncertainty. Ruijs et al. (2013, 2015) empirically check the transformation function between marketable outputs and environmental goods by parametric estimation and confirm a competitive relationship. Bostian and Herlihy (2014) expect that agricultural production contributes to the degradation of wetland conditions due to drainage, channelization and run-off, but qualify this by claiming that the biophysical relationship is not exactly known.

The assumed competitive relationship has been contested in recent literature. Several contributions argue that some environmental goods are complementary to conventional production for lower levels of the environmental good and competitive for higher levels (Hodge 2008). Such a complementary-competitive relationship is hypothesized for *inter alia* the environmental quality of grassland and livestock production (Vatn 2002), pollinator habitat

and crop production (Wossink and Swinton 2007), and the entire ecosystem on the farm and total agricultural production (Hodge 2000).

There is nonetheless only limited empirical evidence of this relationship. Peerlings (2004) arrives at a competitive relationship between milk production on the one hand, and wildlife and landscape services on the other hand. Havlik (2005) finds evidence of a complementary-competitive relationship between grassland biodiversity and cattle production. Sauer and Wossink (2013) approximate a 'bundled' environmental good as the total green payments provided by the CAP. They apply a flexible transformation function and obtain a complementary relationship for most farms and a competitive relationship for a minority of farms.

5.3 The Trade-Off Between Inputs and Environmental Goods

Using the envelope theorem, the trade-off between inputs and environmental goods can also be inferred (Chambers et al. 1996, 1998):

$$-\frac{\partial_x \vec{D}_\Psi(x, y, e; g_x, g_y, g_e)}{\partial_e \vec{D}_\Psi(x, y, e; g_x, g_y, g_e)} = \frac{w}{u} \geq 0 \tag{58}$$

By treating an environmental good as a conventional output, it is implicitly assumed that the provision of *any* environmental good is non-decreasing for increases in *any* input. Figure 11 shows the production possibility frontier for one input x_1 and one environmental good e_1, holding other inputs, other environmental goods and outputs constant. Equation (58) can in principle be used to compute the shadow value u. However, as most studies focus on the trade-off between environmental goods and marketable outputs, Eq. (58) has not been of interest in practice.

The augmented production economics approach

Färe et al. (2001b), Areal et al. (2012), and Sipilainen and Huhtala (2012) augment a conventional production economics framework (with marketable inputs and outputs) with, respectively, the characteristics of public land conservation (the number of conservation sites, the area at each site and the total area available for fishing), the share of grassland and the Shannon index for crop diversification. This '*augmented production economics approach*' is intuitive for economists as it is an extension of familiar neoclassical models. Interestingly, none of these studies elaborates on the implicit assumption

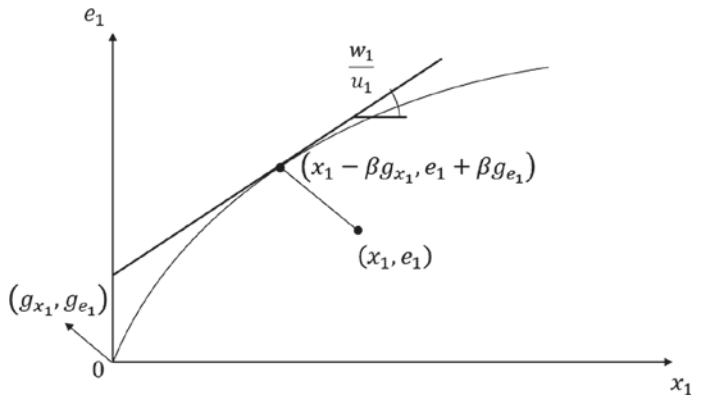

Fig. 11 Trade-off between one input x_1 and one environmental good e_1, holding other inputs, environmental goods and outputs constant

that the provision of an environmental good is non-decreasing in the inputs if one models an environmental good as a conventional output.

One may argue that this assumption holds for inputs that compete for environmental goods jointly produced with marketable outputs. For example, farmers may set aside land and other inputs to produce conservation buffers and cover crops. The same inputs could also have been used to produce marketable outputs (Wossink and Swinton 2007). However, as both inputs and environmental goods are heterogeneous, we argue that the expected relationship between input use and provision of environmental goods may also be non-positive or unclear. We expect a non-positive relationship for inputs that contain environmentally damaging substances. Fertilizer use may lead to nitrogen leaching in the soil and eventually to reduced groundwater quality. It may also volatilize into nitrogen oxide, a GHG (Reinhard et al. 1999). Pesticide use is expected to have a negative impact on farm biodiversity as it suppresses beneficial organisms such as beetles and birds (Skevas et al. 2012). The relationship may depend on the environmental good. For instance, although fertilizer use may decrease groundwater quality, its impact on farm biodiversity is uncertain.

Augmented production economics approaches have focused on the output distance function, holding inputs constant. This may be the reason why the implicit assumption of the non-negative relationship between inputs and environmental goods has not been motivated. Nevertheless, the underlying production technology still depends on inputs. An incorrect assumption about the relationship between inputs and environmental goods also leads to an incorrect computation of the output distance function. Unfortunately, making such an a priori assumption is no trivial task.

The biophysical approach

Several studies veer from the augmented production economics approach. The '*biophysical approach*' considers marketable outputs and environmental goods, but no marketable inputs. Although environmental, non-marketable inputs are generally chosen more sparingly and carefully than in the augmented production economics approach, this is necessarily done on an ad hoc basis, which compromises economic intuition. Macpherson et al. (2010) consider four environmental inputs (percentage edge forest, percentage of impervious surface, percentage of riparian agriculture and road density). Explicitly stating that 'this model specification lacks the clarity of the input–output relationship of a typical model in production economics' (p. 1921), they conduct a correlation analysis with the outputs (per capita income, population density, percentage of wetland and percentage of interior forest) as a robustness test and only partly confirm a positive relationship. Bostian and Herlihy (2014) solely implement joint land use as an input. Bellenger and Herlihy (2009, 2010) and Ruijs et al. (2013) do not even consider any inputs.

Finally, following the previous discussion in the case of multi-equation modelling, a factorially determined production system can also be estimated, where one sub-technology describes the production of abatement outputs.

6 Data Requirements

Assessing pollution-adjusted efficiency requires data on the environmental outputs. One can distinguish between observable or measurable outputs and estimated (approximated) outputs. Point source pollution and non-point source pollution are the main classes of measurable outputs. Point source pollutants can be easily traced back to their source. Examples can be found for wastewater treatment plants, factories for which pollution is discarded through pipes, and to some extent animals in agriculture. In the benchmarking literature, one can find applications in various sectors such as power plants, or the petroleum, cement or pulp and paper industry (Coggins and Swinton 1996; Färe et al. 2001a; Kumar Mandal and Madheswaran 2010; Sueyoshi and Goto 2012c, 2015a; Yu et al. 2016; Bostian et al. 2018; Sueyoshi and Wang 2018).[19] Non-point source pollution has several sources and requires monitoring of prohibitive costs. A common example is runoff

[19]See Zhou et al. (2008b) for more applications.

and leaching of agricultural chemical inputs. Another feature of this type of pollution is its variability, since it can be affected by environmental factors such as weather variables and soil quality. In the agricultural sector, a common approach to evaluate environmental impacts like GHGs emissions is life cycle assessment (Ekvall and Finnveden 2001; Finnveden et al. 2009). This method attributes an emission factor (for instance IPCC guidelines) to each agricultural input for the computation of GHGs emissions. Applications in the agricultural sector can be found in Iribarren et al. (2011), Shortall and Barnes (2013), Toma et al. (2013), and Dakpo et al. (2017a).

Since they are related to a physical process, all these data must satisfy the laws of thermodynamics. However, it is very hard to survey all the variables that intervene in the production process, for instance, the oxygen from the atmosphere. It is worth stressing one of our main conclusions, i.e. what matters is an adequate modelling of pollution-generating technologies and not an explicit accounting of the materials balance principle. When the technology is properly modelled, the materials balance principle is implicitly satisfied.

7 Conclusion

This chapter focused on the modelling of pollution-adjusted technologies in a production framework. The first part has been devoted to the appropriate modelling of environmental bads. Earlier contributions suggest that pollutants should be modelled as conventional inputs, as these are assumed to be complements of marketable outputs (e.g. Baumol and Oates 1988; Reinhard et al. 1999, 2000; Hailu and Veeman 2001). However, this approach unrealistically assumes that fixed amounts of inputs can produce an unlimited amount of pollutants (Färe and Grosskopf 2003). Most authors therefore treat pollutants as weakly disposable outputs which have complementary characteristics for lower levels of pollution and competitive characteristics for higher levels of pollution (e.g. Pittman 1983; Färe et al. 2005, 2014). Hundreds of papers assume treat pollutants as weakly disposable outputs (Dakpo et al. 2017b). The rationale is that conventional production increases with pollution, but that clean-up opportunity costs arise for higher pollution levels. This implies that the shadow price of pollution can also become negative. This has been contested by Hailu and Veeman (2001), who therefore model pollution as an input. Färe and Grosskopf (2003) claim that this is a conflation of the choice of the production technology and the directional vector. They put forward that although pollution should

be modelled as a weakly disposable output, it is still possible to choose a directional vector that points towards the complementary part of the frontier, which would result in positive shadow prices.

Modelling pollutants as inputs or weakly disposable outputs are essentially a black-box approach. This may result in unacceptable implications for trade-offs among inputs, outputs and pollutants (Førsund 2009). Coelli et al. (2007) introduce an environmental efficiency measure that complies with the materials balance principle. Instead of adding pollution as an additional variable, polluting inputs and outputs are chosen in a pollution-minimizing way. Murty et al. (2012) model the polluting technology as the intersection of an intended-output technology and a residual-generating technology. This approach of accurately accounting for multiple processes is typical for engineering science and is appealing for economists. The multi-equation approach opens the black box by making the technical relationships between all inputs and outputs explicit. This increase in accuracy does, however, require appropriate knowledge of the production system (what products are produced by the same process, what are the specific inputs and is the role of the other inputs, are products joint or substitutes, etc.). The benefit of the multiple equation representation is that you can model the technology without explicitly introducing the mass balance identity equation and without violating it. However, this approach still requires inherently subjective judgement. It is an open question whether residuals should be treated as outcomes and iso-environmental lines should be used, or whether a structural representation with multiple equations is needed.

While the literature on environmental bads is abundant, interest in environmental goods has only recently emerged. Environmental goods are now commonly modelled as conventional outputs in a distance function framework, analogous to how pollutants have been modelled as conventional inputs in the earlier environmental economics literature. This assumes that there is a competitive relationship with marketable outputs for all levels of the environmental good and that the shadow price of an environmental good is positive. Recent studies have put forward that an environmental good may also be complementary to marketable outputs. Treating an environmental good as a conventional output also implies that its provision is assumed to be non-decreasing in the inputs. For the augmented production economics approach, which includes all marketable inputs, this assumption may be incorrect for at least some inputs. Clearly, inputs such as fertilizers and pesticides decrease the provision of some environmental goods. This critique is somewhat analogous to Murty et al. (2012), who argue that treating a pollutant as an input incorrectly implies that the trade-off between a

pollution-generating input and a pollutant is assumed to be non-positive. One could adapt a biophysical approach and focus on environmental, non-marketable inputs. However, inputs are in that case chosen ad hoc, which compromises economic intuition. An additional difficulty is that environmental goods are considerably more heterogeneous than environmental bads, where the axiomatic properties are better understood.

A potentially complementary-competitive relationship between environmental goods and marketable outputs calls into question whether the WDA should be invoked for an environmental good (Van Huylenbroeck et al. 2007), as has been frequently done for pollutants. This would imply that the shadow price of an environmental good could be negative or positive (Wossink and Swinton 2007). However, the empirical evidence of a complementary-competitive relationship is limited. Moreover, the sheer heterogeneity of inputs and environmental goods complicates the a priori assumption about the trade-off between inputs and environmental goods. The research on modelling environmental goods would benefit from the lessons in modelling environmental bads. Ultimately, we believe that the appropriate modelling of environmental goods requires a multi-equation approach that makes the technical relationships between all inputs and outputs explicit.

We have several paths for future research. First, it should take into account other challenges that have appeared over time: the treatment of abatement outputs (end-of-pipe technologies, adoption of cleaner technologies) and the shadow pricing of undesirable outputs, especially in the multi-equation scenario. Second, while the main focus of this chapter is on material-based technologies, more research is needed for non-materials technologies like the bank industry. If physical laws are not present, does this mean that we can do whatever we want?

References

Adler, N., and N. Volta. 2016. Accounting for externalities and disposability: A directional economic environmental distance function. *European Journal of Operational Research* 250: 314–327.

Aldanondo-Ochoa, A.M., V.L. Casasnovas-Oliva, and M.C. Almansa-Saez. 2017. Cross-constrained measuring the cost-environment efficiency in material balance based frontier models. *Ecological Economics* 142: 46–55.

Areal, F.J., R. Tiffin, and K.G. Balcombe. 2012. Provision of environmental output within a multi-output distance function approach. *Ecological Economics* 78: 47–54.

Arjomandi, A., K.H. Dakpo, and J.H. Seufert. 2018. Have Asian airlines caught up with European Airlines? A by-production efficiency analysis. *Transportation Research Part A: Policy and Practice* 116: 389–403.

Arocena, P., and C. Waddams Price. 2002. Generating efficiency: Economic and environmental regulation of public and private electricity generators in Spain. *International Journal of Industrial Organization* 20: 41–69.

Ayres, R.U. 1998. Eco-thermodynamics: Economics and the second law. *Ecological Economics* 26: 189–209.

Ayres, R.U., and L. Ayres. 2002. *A handbook of industrial ecology*. Northampton, MA: Edward Elgar.

Ayres, R.U., and A.V. Kneese. 1969. Production, consumption, and externalities. *The American Economic Review* 59: 282–297.

Barros, C.P., S. Managi, and R. Matousek. 2012. The technical efficiency of the Japanese banks: Non-radial directional performance measurement with undesirable output. *Omega* 40: 1–8.

Baumgärtner, S. 2004. The Inada conditions for material resource inputs reconsidered. *Environmental & Resource Economics* 29: 307–322.

Baumgärtner, S., H. Dyckhoff, M. Faber, J. Proops, and J. Schiller. 2001. The concept of joint production and ecological economics. *Ecological Economics* 36: 365–372.

Baumol, W.J., and W.E. Oates. 1988. *The theory of environmental policy*. New York: Cambridge University Press.

Bellenger, M.J., and A.T. Herlihy. 2009. An economic approach to environmental indices. *Ecological Economics* 68: 2216–2223.

Bellenger, M.J., and A.T. Herlihy. 2010. Performance-based environmental index weights: Are all metrics created equal? *Ecological Economics* 69: 1043–1050.

Berg, S.A., F.R. Førsund, and E.S. Jansen. 1992. Malmquist indices of productivity growth during the deregulation of Norwegian banking, 1980–89. *The Scandinavian Journal of Economics* 94: S211–S228.

Bi, G.B., Y.Y. Shao, W. Song, F. Yang, and Y. Luo. 2018. A performance evaluation of China's coal-fired power generation with pollutant mitigation options. *Journal of Cleaner Production* 171: 867–876.

Bostian, M.B., and A.T. Herlihy. 2014. Valuing tradeoffs between agricultural production and wetland condition in the U.S. Mid-Atlantic region. *Ecological Economics* 105: 284–291.

Bostian, M., R. Färe, S. Grosskopf, T. Lundgren, and W.L. Weber. 2018. Time substitution for environmental performance: The case of Swedish manufacturing. *Empirical Economics* 54: 129–152.

Boyd, G.A., G. Tolley, and J. Pang. 2002. Plant level productivity, efficiency, and environmental performance of the container glass industry. *Environmental and Resource Economics* 23: 29–43.

Callan, S.J., and J.M. Thomas. 2009. *Environmental economics and management: Theory, policy and applications*. Mason: Cengage Learning.

Chambers, R.G. 1988. *Applied production analysis: A dual approach*. Cambridge: Cambridge University Press.

Chambers, R.G., Y. Chung, and R. Färe. 1996. Benefit and distance functions. *Journal of Economic Theory* 70: 407–419.

Chambers, R.G., Y. Chung, and R. Färe. 1998. Profit, directional distance functions, and Nerlovian efficiency. *Journal of Optimization Theory and Applications* 98: 351–364.

Chambers, R.G., T. Serra, and A. Oude Lansink. 2014. On the pricing of undesirable state-contingent outputs. *European Review of Agricultural Economics* 41: 485–509.

Chang, T.-P., and J.-L. Hu. 2010. Total-factor energy productivity growth, technical progress, and efficiency change: An empirical study of China. *Applied Energy* 87: 3262–3270.

Chen, C.-M. 2013. Evaluating eco-efficiency with data envelopment analysis: An analytical reexamination. *Annals of Operations Research* 214: 49–71.

Chen, C.-M., and M.A. Delmas. 2012. Measuring eco-inefficiency: A new frontier approach. *Operations Research* 60: 1064–1079.

Chen, P., M. Yu, S. Managi, and C. Chang. 2011. Non-radial directional performance measurement with undesirable outputs. Working Paper, Tohoku University, Japan, 2010.

Chen, L., Y.-M. Wang, and F. Lai. 2017. Semi-disposability of undesirable outputs in data envelopment analysis for environmental assessments. *European Journal of Operational Research* 260: 655–664.

Chung, Y.H., R. Färe, and S. Grosskopf. 1997. Productivity and undesirable outputs: A directional distance function approach. *Journal of Environmental Management* 51: 229–240.

Coelli, T., L. Lauwers, and G. Van Huylenbroeck. 2005. Formulation of technical, economic and environmental efficiency measures that are consistent with the materials balance condition. School of Economics, University of Queensland, Australia.

Coelli, T., L. Lauwers, and G. Van Huylenbroeck. 2007. Environmental efficiency measurement and the materials balance condition. *Journal of Productivity Analysis* 28: 3–12.

Coggins, J.S., and J.R. Swinton. 1996. The price of pollution: A dual approach to valuing SO2 allowances. *Journal of Environmental Economics and Management* 30: 58–72.

Considine, T.J., and D.F. Larson. 2006. The environment as a factor of production. *Journal of Environmental Economics and Management* 52: 645–662.

Cooper, W.W., K.S. Park, and J.T. Pastor. 1999a. RAM: A range adjusted measure of inefficiency for use with additive models, and relations to other models and measures in DEA. *Journal of Productivity Analysis* 11: 5–42.

Cooper, W.W., K.S. Park, and G. Yu. 1999b. IDEA and AR-IDEA: Models for dealing with imprecise data in DEA. *Management Science* 45: 597–607.

Courcelle, C., M.-P. Kestemont, D. Tyteca, and M. Installé. 1998. Assessing the economic and environmental performance of municipal solid waste collection and sorting programmes. *Waste Management & Research* 16: 253–262.

Cui, Q., and Y. Li. 2016. Airline energy efficiency measures considering carbon abatement: A new strategic framework. *Transportation Research Part D-Transport and Environment* 49: 246–258.

Cui, Q., and Y. Li. 2017a. Airline efficiency measures under CNG2020 strategy: An application of a dynamic by-production model. *Transportation Research Part A-Policy and Practice* 106: 130–143.

Cui, Q., and Y. Li. 2017b. Airline environmental efficiency measures considering materials balance principles: An application of a network range-adjusted measure with weak-G disposability. *Journal of Environmental Planning and Management* 61 (13): 1–21.

Cui, Q., and Y. Li. 2018. Airline dynamic efficiency measures with a dynamic RAM with unified natural & managerial disposability. *Energy Economics* 75: 534–546.

Dakpo, K.H., P. Jeanneaux, and L. Latruffe. 2017a. Greenhouse gas emissions and efficiency in French sheep meat farming: A non-parametric framework of pollution-adjusted technologies. *European Review of Agricultural Economics* 44: 33–65.

Dakpo, K.H., P. Jeanneaux, and L. Latruffe. 2017b. Modelling pollution-generating technologies in performance benchmarking: Recent developments, limits and future prospects in the nonparametric framework. *European Journal of Operational Research* 250: 347–359.

De Bruyn, S. 2002. Dematerialization and rematerialization as two recurring phenomena of industrial ecology. In *A handbook of industrial ecology*, ed. R.U. Ayres and L. Ayres, 209–222. Northampton, MA: Edward Elgar.

De Koeijer, T.J., G.A. Wossink, P.C. Struik, and J.A. Renkema. 2002. Measuring agricultural sustainability in terms of efficiency: The case of Dutch sugar beet growers. *Journal of Environmental Economics and Management* 66: 9–17.

Domazlicky, B.R., and W.L. Weber. 2004. Does environmental protection lead to slower productivity growth in the chemical industry? *Environmental and Resource Economics* 28: 301–324.

Du, K., H. Lu, and K. Yu. 2013. Sources of the potential CO_2 emission reduction in China: A nonparametric metafrontier approach. *Applied Energy* 115: 491–501.

Dyckhoff, H., and K. Allen. 2001. Measuring ecological efficiency with data envelopment analysis (DEA). *European Journal of Operational Research* 132: 312–325.

Ebert, U., and H. Welsch. 2007. Environmental emissions and production economics: Implications of the materials balance. *American Journal of Agricultural Economics* 89: 287–293.

Ekvall, T., and G. Finnveden. 2001. Allocation in ISO 14041—A critical review. *Journal of Cleaner Production* 9: 197–208.

Färe, R., and S. Grosskopf. 1996. *Intertemporal production frontiers: With dynamic DEA*. Norwell, MA: Springer.

Färe, R., and S. Grosskopf. 2003. Nonparametric productivity analysis with undesirable outputs: Comment. *American Journal of Agricultural Economics* 85: 1070–1074.

Färe, R., and S. Grosskopf. 2004a. Modeling undesirable factors in efficiency evaluation: Comment. *European Journal of Operational Research* 157: 242–245.

Färe, R., and S. Grosskopf. 2004b. *New directions: Efficiency and productivity*. New York: Springer Science + Business Media.

Färe, R., and S. Grosskopf. 2005. *New directions: Efficiency and productivity*. New York: Springer.

Färe, R., S. Grosskopf, and C. Pasurka. 1986. Effects on relative efficiency in electric power generation due to environmental controls. *Resources and Energy* 8: 167–184.

Färe, R., S. Grosskopf, C.K. Lovell, and C. Pasurka. 1989. Multilateral productivity comparisons when some outputs are undesirable: A nonparametric approach. *The Review of Economics and Statistics* 71: 90–98.

Färe, R., S. Grosskopf, and D. Tyteca. 1996. An activity analysis model of the environmental performance of firms—Application to fossil-fuel-fired electric utilities. *Ecological Economics* 18: 161–175.

Färe, R., S. Grosskopf, and J.C.A. Pasurka. 2001a. Accounting for air pollution emissions in measures of state manufacturing productivity growth. *Journal of Regional Science* 41: 381–409.

Färe, R., S. Grosskopf, and W.L. Weber. 2001b. Shadow prices of missouri public conservation land. *Public Finance Review* 29: 444–460.

Färe, R., S. Grosskopf, D.-W. Noh, and W. Weber. 2005. Characteristics of a polluting technology: Theory and practice. *Journal of Econometrics* 126: 469–492.

Färe, R., S. Grosskopf, and C.A. Pasurka Jr. 2007. Environmental production functions and environmental directional distance functions. *Energy* 32: 1055–1066.

Färe, R., S. Grosskopf, T. Lundgren, P.-O. Marklund, and W. Zhou. 2012. Productivity: Should we include bads? CERE Center for Environmental and Resource Economics.

Färe, R., S. Grosskopf, and C. Pasurka. 2013. Joint production of good and bad outputs with a network application. *Encyclopedia of Energy, Natural Resources and Environmental Economics* 2: 109–118.

Färe, R., S. Grosskopf, and C.A. Pasurka. 2014. Potential gains from trading bad outputs: The case of U.S. electric power plants. *Resource and Energy Economics* 36: 99–112.

Faucheux, S. 1994. Energy analysis and sustainable development. In *Valuing the environment: Methodological and measurement issues*, ed. R. Pethig, 325–346. Dordrecht: Springer.

Fernández, C., G. Koop, and M.F.J. Steel. 2002. Multiple-output production with undesirable outputs: An application to nitrogen surplus in agriculture. *Journal of the American Statistical Association* 97: 432–442.

Fernández, C., G. Koop, and M.F. Steel. 2005. Alternative efficiency measures for multiple-output production. *Journal of Econometrics* 126: 411–444.

Finnveden, G., M.Z. Hauschild, T. Ekvall, J. Guinee, R. Heijungs, S. Hellweg, A. Koehler, D. Pennington, and S. Suh. 2009. Recent developments in life cycle assessment. *Journal of Environmental Management* 91: 1–21.

Fleishman, R., R. Alexander, S. Bretschneider, and D. Popp. 2009. Does regulation stimulate productivity? The effect of air quality policies on the efficiency of US power plants. *Energy Policy* 37: 4574–4582.

Førsund, F.R. 1972. Allocation in space and environmental pollution. *The Swedish Journal of Economics* 74: 19–34.

Førsund, F.R. 1973. Externalities, environmental pollution and allocation in space: A general equilibrium approach. *Regional and Urban Economics* 3: 3–32.

Førsund, F.R. 1998. Pollution modelling and multiple-output production theory. Department of Economics, University of Oslo.

Førsund, F.R. 2009. Good modelling of bad outputs: Pollution and multiple-output production. *International Review of Environmental and Resource Economics* 3: 1–38.

Førsund, F.R. 2017. Multi-equation modelling of desirable and undesirable outputs satisfying the materials balance. *Empirical Economics* 54: 67–99.

Førsund, F.R. 2018. Pollution meets efficiency: Multi-equation modelling of generation of pollution and related efficiency measures. In *Energy, environment and transitional green growth in China*, ed. R. Pang, X. Bai, and K. Lovell, 37–79. Springer: Singapore.

Frisch, R. 1965. *Theory of production*. Dordrecht: Reidel Publishing Company.

Frondel, M., J. Horbach, and K. Rennings. 2007. End-of-pipe or cleaner production? An empirical comparison of environmental innovation decisions across OECD countries. *Business Strategy and the Environment* 16: 571–584.

Georgescu-Roegen, N. 1971. *The entropy law and the economic process*. Cambridge, MA: Harvard University Press.

Glavič, P., and R. Lukman. 2007. Review of sustainability terms and their definitions. *Journal of Cleaner Production* 15: 1875–1885.

Guesmi, B., and T. Serra. 2015. Can we improve farm performance? The determinants of farm technical and environmental efficiency. *Applied Economic Perspectives and Policy* 37: 692–717.

Hailu, A., and T.S. Veeman. 2000. Environmentally sensitive productivity analysis of the Canadian pulp and paper industry, 1959–1994: An input distance function approach. *Journal of Environmental Economics and Management* 40: 251–274.

Hailu, A., and T.S. Veeman. 2001. Non-parametric productivity analysis with undesirable outputs: An application to the Canadian pulp and paper industry. *American Journal of Agricultural Economics* 83: 605–616.

Hammar, H., and Å. Löfgren. 2010. Explaining adoption of end of pipe solutions and clean technologies—Determinants of firms' investments for reducing emissions to air in four sectors in Sweden. *Energy Policy* 38: 3644–3651.

Hampf, B. 2013. Separating environmental efficiency into production and abatement efficiency: A nonparametric model with application to US power plants. *Journal of Productivity Analysis* 41: 457–473.

Hampf, B. 2015. Estimating the materials balance condition: A Stochastic frontier approach. Darmstadt Discussion Papers in Economics, 226.

Hampf, B. 2018a. Cost and environmental efficiency of U.S. electricity generation: Accounting for heterogeneous inputs and transportation costs. *Energy* 163: 932–941.

Hampf, B. 2018b. Measuring inefficiency in the presence of bad outputs: Does the disposability assumption matter? *Empirical Economics* 54: 101–127.

Hampf, B., and K.L. Rødseth. 2015. Carbon dioxide emission standards for U.S. power plants: An efficiency analysis perspective. *Energy Economics* 50: 140–153.

Hampf, B., and K.L. Rødseth. 2017. Optimal profits under environmental regulation: The benefits from emission intensity averaging. *Annals of Operations Research* 255: 367–390.

Havlik, P. 2005. Joint production under uncertainty and multifunctionality of agriculture: Policy considerations and applied analysis. *European Review of Agricultural Economics* 32: 489–515.

Hellweg, S., G. Doka, G. Finnveden, and K. Hungerbühler. 2005. Assessing the eco-efficiency of end-of-pipe technologies with the environmental cost efficiency indicator. *Journal of Industrial Ecology* 9: 189–203.

Hoang, V.-N., and M. Alauddin. 2011. Input-orientated data envelopment analysis framework for measuring and decomposing economic, environmental and ecological efficiency: An application to OECD agriculture. *Environmental and Resource Economics* 51: 431–452.

Hoang, V.-N., and T. Coelli. 2011. Measurement of agricultural total factor productivity growth incorporating environmental factors: A nutrients balance approach. *Journal of Environmental Economics and Management* 62: 462–474.

Hoang, V.-N., and T.T. Nguyen. 2013. Analysis of environmental efficiency variations: A nutrient balance approach. *Ecological Economics* 86: 37–46.

Hoang, V.-N., and D.S.P. Rao. 2010. Measuring and decomposing sustainable efficiency in agricultural production: A cumulative exergy balance approach. *Ecological Economics* 69: 1765–1776.

Hodge, I.D. 2000. Agri-environmental relationships and the choice of policy mechanism. *The World Economy* 23: 257–273.

Hodge, I.D. 2008. *To what extent are environmental externalities a joint product of agriculture?* OECD Publishing.

Hua, Z., Y. Bian, and L. Liang. 2007. Eco-efficiency analysis of paper mills along the Huai River: An extended DEA approach. *Omega* 35: 578–587.

Iribarren, D., A. Hospido, M.T. Moreira, and G. Feijoo. 2011. Benchmarking environmental and operational parameters through eco-efficiency criteria for dairy farms. *Science of the Total Environment* 409: 1786–1798.

Jahanshahloo, G.R., A. Hadi Vencheh, A.A. Foroughi, and R. Kazemi Matin. 2004. Inputs/outputs estimation in DEA when some factors are undesirable. *Applied Mathematics and Computation* 156: 19–32.

James, D. 1985. Environmental economics, industrial process models, and regional-residuals management models. In *Handbook of natural resource and energy economics*, Chapter 7, 271–324. Amsterdam: Elsevier.

Kao, C., and S.-N. Hwang. 2017. Efficiency evaluation in the presence of undesirable outputs: The most favorable shadow price approach. *Annals of Operations Research* 278: 1–12.

Keeble, B.R. 1988. The Brundtland report: 'Our common future'. *Medicine and War* 4: 17–25.

Kneese, A.V., R.U. Ayres, and R.C.D'Arge. 1970. *Economics and the environment: A materials balance approach*. Washington, DC: Resources for the Future.

Kortelainen, M., and T. Kuosmanen. 2004. Measuring eco-efficiency of production: A frontier approach, EconWPA Working Paper No. 0411004. Department of Economics, Washington University St. Louis, MO.

Kumar, S. 2006. Environmentally sensitive productivity growth: A global analysis using Malmquist-Luenberger index. *Ecological Economics* 56: 280–293.

Kumar, S., and S. Managi. 2011. Non-separability and substitutability among water pollutants: Evidence from India. *Environment and Development Economics* 16: 709–733.

Kumar Mandal, S., and S. Madheswaran. 2010. Environmental efficiency of the Indian cement industry: An interstate analysis. *Energy Policy* 38: 1108–1118.

Kumbhakar, S.C., and E.G. Tsionas. 2016. The good, the bad and the technology: Endogeneity in environmental production models. *Journal of Econometrics* 190: 315–327.

Kuosmanen, T. 2005. Weak disposability in nonparametric production analysis with undesirable outputs. *American Journal of Agricultural Economics* 87: 1077–1082.

Kuosmanen, N. 2014. Estimating stocks and flows of nitrogen: Application of dynamic nutrient balance to European agriculture. *Ecological Economics* 108: 68–78.

Kuosmanen, T., and R. Kazemi Matin. 2011. Duality of weakly disposable technology. *Omega* 39: 504–512.

Kuosmanen, T., and M. Kortelainen. 2005. Measuring Eco-efficiency of production with data envelopment analysis. *Journal of Industrial Ecology* 9: 59–72.

Kuosmanen, N., and T. Kuosmanen. 2013. Modeling cumulative effects of nutrient surpluses in agriculture: A dynamic approach to material balance accounting. *Ecological Economics* 90: 159–167.

Kuosmanen, T., and V. Podinovski. 2009. Weak disposability in nonparametric production analysis: Reply to Färe and Grosskopf. *American Journal of Agricultural Economics* 91: 539–545.

Lauwers, L. 2009. Justifying the incorporation of the materials balance principle into frontier-based eco-efficiency models. *Ecological Economics* 68: 1605–1614.

Lauwers, L., G. Van Huylenbroeck, and G. Rogiers. 1999. Technical, economic and environmental efficiency analysis of pig fattening farms. 9th European Congress of Agricultural Economists. Warsaw, Poland.

Lee, J.-D., J.-B. Park, and T.-Y. Kim. 2002. Estimation of the shadow prices of pollutants with production/environment inefficiency taken into account: A nonparametric directional distance function approach. *Journal of Environmental Management* 64: 365–375.

Lee, B.L., C. Wilson, C.A. Pasurka, H. Fujii, and S. Managi. 2016. Sources of airline productivity from carbon emissions: An analysis of operational performance under good and bad outputs. *Journal of Productivity Analysis* 47: 223–246.

Leleu, H. 2013. Shadow pricing of undesirable outputs in nonparametric analysis. *European Journal of Operational Research* 231: 474–480.

Levkoff, S.B. 2013. Efficiency trends in US coal-fired energy production & the 1990 Clean Air Act amendment: A nonparametric approach. Department of Economics, University of California, San Diego, USA.

Liang, L., Y. Li, and S. Li. 2009. Increasing the discriminatory power of DEA in the presence of the undesirable outputs and large dimensionality of data sets with PCA. *Expert Systems with Applications* 36: 5895–5899.

Liu, W., and J. Sharp. 1999. DEA models via goal programming. In *Data envelopment analysis in the service Sector*, 79–101. Wiesbaden: Springer.

Lovell, C.A.K., J.T. Pastor, and J.A. Turner. 1995. Measuring macroeconomic performance in the OECD: A comparison of European and non-European countries. *European Journal of Operational Research* 87: 507–518.

Lozano, S. 2015. A joint-inputs network DEA approach to production and pollution-generating technologies. *Expert Systems with Applications* 42: 7960–7968.

Lozano, S., and E. Gutiérrez. 2008. Non-parametric frontier approach to modelling the relationships among population, GDP, energy consumption and CO_2 emissions. *Ecological Economics* 66: 687–699.

Lozano, S., and E. Gutiérrez. 2011. Slacks-based measure of efficiency of airports with airplanes delays as undesirable outputs. *Computers & Operations Research* 38: 131–139.

Macpherson, A.J., P.P. Principe, and E.R. Smith. 2010. A directional distance function approach to regional environmental–economic assessments. *Ecological Economics* 69: 1918–1925.

Mahlberg, B., and B.K. Sahoo. 2011. Radial and non-radial decompositions of Luenberger productivity indicator with an illustrative application. *International Journal of Production Economics* 131: 721–726.

Mahlberg, B., M. Luptacik, and B.K. Sahoo. 2011. Examining the drivers of total factor productivity change with an illustrative example of 14 EU countries. *Ecological Economics* 72: 60–69.

Malikov, E., S. Kumbhakar, and E. Tsionas. 2015. Bayesian approach to disentangling technical and environmental productivity. *Econometrics* 3: 443–465.

Malikov, E., R. Bokusheva, and S.C. Kumbhakar. 2018. A hedonic-output-index-based approach to modeling polluting technologies. *Empirical Economics* 54: 287–308.

Manello, A. 2012. Efficiency and productivity analysis in presence of undesirable output: An extended literature review. University of Bergamo-Faculty of Engineering, 127pp. http://aisberg.unibg.it/bitstream/10446/26695/1/A.Manello%20-%20PhD%20thesis.pdf.

Marklund, P.-O., and E. Samakovlis. 2007. What is driving the EU burden-sharing agreement: Efficiency or equity? *Journal of Environmental Management* 85: 317–329.

Mehta, L. 2013. *The limits to scarcity: Contesting the politics of allocation*. London and Washington, DC: Earthscan.

Murty, S. 2010a. Externalities and fundamental nonconvexities: A reconciliation of approaches to general equilibrium externality modeling and implications for decentralization. *Journal of Economic Theory* 145: 331–353.

Murty, S. 2010b. On the theory of a firm: The case of by-production of emissions. *Warwick Economic Research Papers* 934: 1–45.

Murty, S. 2015. On the properties of an emission-generating technology and its parametric representation. *Economic Theory* 60: 243–282.

Murty, S., and Russell, R.R. 2002. On modeling pollution generating technologies. Department of Economics, University of California, Riverside.

Murty, S., and R.R. Russell. 2016. Modeling emission-generating technologies: Reconciliation of axiomatic and by-production approaches. *Empirical Economics* 54: 7–30.

Murty, S., R. Robert Russell, and S.B. Levkoff. 2012. On modeling pollution-generating technologies. *Journal of Environmental Economics and Management* 64: 117–135.

Olajire, A.A. 2010. CO_2 capture and separation technologies for end-of-pipe applications—A review. *Energy* 35: 2610–2628.

Paul, C.J.M., V.E. Ball, R.G. Felthoven, A. Grube, and R.F. Nehring. 2002. Effective costs and chemical use in United States agricultural production: Using the environment as a "free" input. *American Journal of Agricultural Economics* 84: 902–915.

Peerlings, J. 2004. Wildlife and landscape services production in Dutch dairy farming; jointness and transaction costs. *European Review of Agriculture Economics* 31: 427–449.

Pérez, K., M.C. González-Araya, and A. Iriarte. 2017. Energy and GHG emission efficiency in the Chilean manufacturing industry: Sectoral and regional analysis by DEA and Malmquist indexes. *Energy Economics* 66: 290–302.

Pethig, R. 2003. The 'materials balance approach' to pollution: Its origin, implications and acceptance. Economics Discussion Paper no. 105-03, University of Siegen.

Pethig, R. 2006. Non-linear production, abatement, pollution and materials balance reconsidered. *Journal of Environmental Economics and Management* 51: 185–204.

Pham, M.D., and V. Zelenyuk. 2018. Weak disposability in nonparametric production analysis: A new taxonomy of reference technology sets. *European Journal of Operational Research* 274:186–198.

Picazo-Tadeo, A.J., and D. Prior. 2009. Environmental externalities and efficiency measurement. *Journal of Environmental Management* 90: 3332–3339.

Picazo-Tadeo, A.J., E. Reig-Martinez, and F. Hernandez-Sancho. 2005. Directional distance functions and environmental regulation. *Resource and Energy Economics* 27: 131–142.

Pigou, A.C. 1920. *The economics of welfare*. Basingstoke: Macmillan.

Pittman, R.W. 1983. Multilateral productivity comparisons with undesirable outputs. *The Economic Journal* 93: 883–891.

Podinovski, V.V., and T. Kuosmanen. 2011. Modelling weak disposability in data envelopment analysis under relaxed convexity assumptions. *European Journal of Operational Research* 211: 577–585.

Porter, M.E. 1991. America's green strategy. *Scientific American* 264: 168.

Porter, M.E., and C. van der Linde. 1995. Toward a new conception of the environment-competitiveness relationship. *Journal of Economic Perspectives* 9: 97–118.

Prior, D. 2006. Efficiency and total quality management in health care organizations: A dynamic frontier approach. *Annals of Operations Research* 145: 281–299.

Ray, S.C., K. Mukherjee, and A. Venkatesh. 2017. Nonparametric measures of efficiency in the presence of undesirable outputs: A by-production approach. *Empirical Economics* 54: 31–65.

Reinhard, S., C.A.K. Lovell, and G. Thijssen. 1999. Econometric estimation of technical and environmental efficiency: An application to Dutch dairy farms. *American Journal of Agricultural Economics* 81: 44–60.

Reinhard, S., C.A.K. Lovell, and G.J. Thijssen. 2000. Environmental efficiency with multiple environmentally detrimental variables; estimated with SFA and DEA. *European Journal of Operational Research* 121: 287–303.

Reinhard, S., C.A.K. Lovell, and G. Thijssen. 2002. Analysis of environmental efficiency variation. *American Journal of Agricultural Economics* 84: 1054–1065.

Rennings, K., A. Ziegler, K. Ankele, and E. Hoffmann. 2006. The influence of different characteristics of the EU environmental management and auditing scheme on technical environmental innovations and economic performance. *Ecological Economics* 57: 45–59.

Rødseth, K.L. 2013. Capturing the least costly way of reducing pollution: A shadow price approach. *Ecological Economics* 92: 16–24.

Rødseth, K.L. 2014. Efficiency measurement when producers control pollutants: A non-parametric approach. *Journal of Productivity Analysis* 42: 211–223.

Rødseth, K.L. 2015. Axioms of a polluting technology: A materials balance approach. *Environmental and Resource Economics* 67: 1–22.

Rødseth, K.L. 2016. Environmental efficiency measurement and the materials balance condition reconsidered. *European Journal of Operational Research* 250: 342–346.

Rødseth, K.L., and E. Romstad. 2013. Environmental regulations, producer responses, and secondary benefits: Carbon dioxide reductions under the acid rain program. *Environmental and Resource Economics* 59: 111–135.

Roshdi, I., M. Hasannasab, D. Margaritis, and P. Rouse. 2018. Generalised weak disposability and efficiency measurement in environmental technologies. *European Journal of Operational Research* 266: 1000–1012.

Ruijs, A., M. Kortelainen, A. Wossink, C.J.E. Schulp, and R. Alkemade. 2015. Opportunity cost estimation of ecosystem services. *Environmental and Resource Economics* 66: 717–747.

Ruijs, A., A. Wossink, M. Kortelainen, R. Alkemade, and C.J.E. Schulp. 2013. Trade-off analysis of ecosystem services in Eastern Europe. *Ecosystem Services* 4: 82–94.

Sahoo, B.K., M. Luptacik, and B. Mahlberg. 2011. Alternative measures of environmental technology structure in DEA: An application. *European Journal of Operational Research* 215: 750–762.

Sarkis, J., and J.J. Cordeiro. 2001. An empirical evaluation of environmental efficiencies and firm performance: Pollution prevention versus end-of-pipe practice. *European Journal of Operational Research* 135: 102–113.

Sauer, J., and A. Wossink. 2013. Marketed outputs and non-marketed ecosystem services: The evaluation of marginal costs. *European Review of Agricultural Economics* 40: 573–603.

Scheel, H. 2001. Undesirable outputs in efficiency valuations. *European Journal of Operational Research* 132: 400–410.

Seiford, L.M., and J. Zhu. 2002. Modeling undesirable factors in efficiency evaluation. *European Journal of Operational Research* 142: 16–20.

Serra, T., R.G. Chambers, and A. Oude Lansink. 2014. Measuring technical and environmental efficiency in a state-contingent technology. *European Journal of Operational Research* 236: 706–717.

Seufert, J.H., A. Arjomandi, and K.H. Dakpo. 2017. Evaluating airline operational performance: A Luenberger-Hicks-Moorsteen productivity indicator. *Transportation Research Part E: Logistics and Transportation Review* 104: 52–68.

Shen, Z., J.-P. Boussemart, and H. Leleu. 2017. Aggregate green productivity growth in OECD's countries. *International Journal of Production Economics* 189: 30–39.

Shephard, R.W. 1953. *Cost and production functions*. DTIC Document.

Shephard, R.W. 1970. *Theory of cost and production functions*. Princeton: Princeton University Press.

Shephard, R.W., and R. Färe. 1974. The law of diminishing returns. *Zeitschrift für Nationalökonomie* 34: 69–90.

Shortall, O., and A. Barnes. 2013. Greenhouse gas emissions and the technical efficiency of dairy farmers. *Ecological Indicators* 29: 478–488.

Sipilainen, T., and A. Huhtala. 2012. Opportunity costs of providing crop diversity in organic and conventional farming: Would targeted environmental policies make economic sense? *European Review of Agricultural Economics* 40: 441–462.

Skevas, T., A.O. Lansink, and S.E. Stefanou. 2012. Measuring technical efficiency in the presence of pesticide spillovers and production uncertainty: The case of Dutch arable farms. *European Journal of Operational Research* 223: 550–559.

Song, M., S. Wang, and W. Liu. 2014. A two-stage DEA approach for environmental efficiency measurement. *Environmental Monitoring and Assessment* 186: 3041–3051.

Song, W., G.-B. Bi, J. Wu, and F. Yang. 2017. What are the effects of different tax policies on China's coal-fired power generation industry? An empirical research from a network slacks-based measure perspective. *Journal of Cleaner Production* 142: 2816–2827.

Sueyoshi, T., and M. Goto. 2010. Should the US clean air act include CO_2 emission control?: Examination by data envelopment analysis. *Energy Policy* 38: 5902–5911.

Sueyoshi, T., and M. Goto. 2011a. DEA approach for unified efficiency measurement: Assessment of Japanese fossil fuel power generation. *Energy Economics* 33: 292–303.

Sueyoshi, T., and M. Goto. 2011b. Measurement of returns to scale and damages to scale for DEA-based operational and environmental assessment: How to manage desirable (good) and undesirable (bad) outputs? *European Journal of Operational Research* 211: 76–89.

Sueyoshi, T., and M. Goto. 2011c. Methodological comparison between two unified (operational and environmental) efficiency measurements for environmental assessment. *European Journal of Operational Research* 210: 684–693.

Sueyoshi, T., and M. Goto. 2012a. Data envelopment analysis for environmental assessment: Comparison between public and private ownership in petroleum industry. *European Journal of Operational Research* 216: 668–678.

Sueyoshi, T., and M. Goto. 2012b. DEA environmental assessment of coal fired power plants: Methodological comparison between radial and non-radial models. *Energy Economics* 34: 1854–1863.

Sueyoshi, T., and M. Goto. 2012c. DEA radial measurement for environmental assessment and planning: Desirable procedures to evaluate fossil fuel power plants. *Energy Policy* 41: 422–432.

Sueyoshi, T., and M. Goto. 2012d. Environmental assessment by DEA radial measurement: US coal-fired power plants in ISO (Independent System Operator) and RTO (Regional Transmission Organization). *Energy Economics* 34: 663–676.

Sueyoshi, T., and M. Goto. 2015a. DEA environmental assessment in time horizon: Radial approach for Malmquist index measurement on petroleum companies. *Energy Economics* 51: 329–345.

Sueyoshi, T., and M. Goto. 2015b. Japanese fuel mix strategy after disaster of Fukushima Daiichi nuclear power plant: Lessons from international comparison among industrial nations measured by DEA environmental assessment in time horizon. *Energy Economics* 52: 87–103.

Sueyoshi, T., and M. Goto. 2018a. Difficulties and remedies on DEA environmental assessment. *Journal of Economic Structures* 7: 2193–2409.

Sueyoshi, T., and M. Goto. 2018b. Resource utilization for sustainability enhancement in Japanese industries. *Applied Energy* 228: 2308–2320.

Sueyoshi, T., and D. Wang. 2014. Radial and non-radial approaches for environmental assessment by Data Envelopment Analysis: Corporate sustainability and effective investment for technology innovation. *Energy Economics* 45: 537–551.

Sueyoshi, T., and D. Wang. 2018. DEA environmental assessment on US petroleum industry: Non-radial approach with translation invariance in time horizon. *Energy Economics* 72: 276–289.

Sueyoshi, T., and Y. Yuan. 2015. Comparison among U.S. industrial sectors by DEA environmental assessment: Equipped with analytical capability to handle zero or negative in production factors. *Energy Economics* 52: 69–86.

Sueyoshi, T., and Y. Yuan. 2016. Returns to damage under undesirable congestion and damages to return under desirable congestion measured by DEA environmental assessment with multiplier restriction: Economic and energy planning for social sustainability in China. *Energy Economics* 56: 288–309.

Sueyoshi, T., M. Goto, and T. Ueno. 2010. Performance analysis of US coal-fired power plants by measuring three DEA efficiencies. *Energy Policy* 38: 1675–1688.

Sueyoshi, T., M. Goto, and M. Sugiyama. 2013. DEA window analysis for environmental assessment in a dynamic time shift: Performance assessment of US coal-fired power plants. *Energy Economics* 40: 845–857.

Sueyoshi, T., M. Goto, and D. Wang. 2017a. Malmquist index measurement for sustainability enhancement in Chinese municipalities and provinces. *Energy Economics* 67: 554–571.

Sueyoshi, T., Y. Yuan, and M. Goto. 2017b. A literature study for DEA applied to energy and environment. *Energy Economics* 62: 104–124.

Sueyoshi, T., A. Li, and Y. Gao. 2018. Sector sustainability on fossil fuel power plants across Chinese provinces: Methodological comparison among radial, non-radial and intermediate approaches under group heterogeneity. *Journal of Cleaner Production* 187: 819–829.

Sun, C., X. Liu, and A. Li. 2018. Measuring unified efficiency of Chinese fossil fuel power plants: Intermediate approach combined with group heterogeneity and window analysis. *Energy Policy* 123: 8–18.

Telle, K., and J. Larsson. 2007. Do environmental regulations hamper productivity growth? How accounting for improvements of plants' environmental performance can change the conclusion. *Ecological Economics* 61: 438–445.

Toma, L., M. March, A. W. Stott, and D.J. Roberts. 2013. Environmental efficiency of alternative dairy systems: A productive efficiency approach. *Journal of Dairy Science* 96: 7014–7031.

Tone, K. 2004. Dealing with undesirable outputs in DEA: A Slacks-Based Measure (SBM) approach. *Nippon Opereshonzu, Risachi Gakkai Shunki Kenkyu Happyokai Abusutorakutoshu* 2004: 44–45.

Tyteca, D. 1996. On the measurement of the environmental performance of firms—A literature review and a productive efficiency perspective. *Journal of Environmental Management* 46: 281–308.

Tyteca, D. 1997. Linear programming models for the measurement of environmental performance of firms—Concepts and empirical results. *Journal of Productivity Analysis* 8: 183–197.

Valadkhani, A., I. Roshdi, and R. Smyth. 2016. A multiplicative environmental DEA approach to measure efficiency changes in the world's major polluters. *Energy Economics* 54: 363–375.

Van Huylenbroeck, G., V. Vandermeulen, E. Mettepenningen, and A. Verspecht. 2007. Multifunctionality of agriculture: A review of definitions, evidence and instruments. *Living Reviews in Landscape Research* 1: 1–43.

Van Meensel, J., L. Lauwers, G. Van Huylenbroeck, and S. Van Passel. 2010. Comparing frontier methods for economic–environmental trade-off analysis. *European Journal of Operational Research* 207: 1027–1040.

van Vuuren, D.P., E. Stehfest, D.E.H.J. Gernaat, M. van den Berg, D.L. Bijl, H.S. de Boer, V. Daioglou, J.C. Doelman, O.Y. Edelenbosch, M. Harmsen, A.F. Hof, and M.A.E. van Sluisveld. 2018. Alternative pathways to the 1.5 °C target reduce the need for negative emission technologies. *Nature Climate Change* 8: 391–397.

Vatn, A. 2002. Multifunctional agriculture: Some consequences for international trade regimes. *European Review of Agricultural Economics* 29: 309–327.

Vencheh, A.H., R. Kazemi Matin, and M. Tavassoli Kajani. 2005. Undesirable factors in efficiency measurement. *Applied Mathematics and Computation* 163: 547–552.

Wang, K., Y.-M. Wei, and X. Zhang. 2012. A comparative analysis of China's regional energy and emission performance: Which is the better way to deal with undesirable outputs? *Energy Policy* 46: 574–584.

Wang, K., Z. Mi, and Y.-M. Wei. 2018a. Will pollution taxes improve joint ecological and economic Efficiency of thermal power industry in China?: A DEA-based materials balance approach. *Journal of Industrial Ecology* 23: 389–401.

Wang, K., Y.-M. Wei, and Z. Huang. 2018b. Environmental efficiency and abatement efficiency measurements of China's thermal power industry: A data envelopment analysis based materials balance approach. *European Journal of Operational Research* 269: 35–50.

Watanabe, M., and K. Tanaka. 2007. Efficiency analysis of Chinese industry: A directional distance function approach. *Energy Policy* 35: 6323–6331.

Weber, W.L., and B. Domazlicky. 2001. Productivity growth and pollution in state manufacturing. *Review of Economics and Statistics* 83: 195–199.

Welch, E., and D. Barnum. 2009. Joint environmental and cost efficiency analysis of electricity generation. *Ecological Economics* 68: 2336–2343.

Wossink, A., and S.M. Swinton. 2007. Jointness in production and farmers' willingness to supply non-marketed ecosystem services. *Ecological Economics* 64: 297–304.

Wu, T.Y., A.W. Mohammad, J.M. Jahim, and N. Anuar. 2010. Pollution control technologies for the treatment of palm oil mill effluent (POME) through end-of-pipe processes. *Journal of Environmental Management* 91: 1467–1490.

Yang, H., and M. Pollitt. 2009. Incorporating both undesirable outputs and uncontrollable variables into DEA: The performance of Chinese coal-fired power plants. *European Journal of Operational Research* 197: 1095–1105.

Yang, H., and M. Pollitt. 2010. The necessity of distinguishing weak and strong disposability among undesirable outputs in DEA: Environmental performance of Chinese coal-fired power plants. *Energy Policy* 38: 4440–4444.

Yörük, B.K., and O. Zaim. 2005. Productivity growth in OECD countries: A comparison with Malmquist indices. *Journal of Comparative Economics* 33: 401–420.

Yu, C., L. Shi, Y.T. Wang, Y. Chang, and B.D. Cheng. 2016. The eco-efficiency of pulp and paper industry in China: An assessment based on slacks-based measure and Malmquist-Luenberger index. *Journal of Cleaner Production* 127: 511–521.

Yu-Ying Lin, E., P.-Y. Chen, and C.-C. Chen. 2013. Measuring the environmental efficiency of countries: A directional distance function metafrontier approach. *Journal of Environmental Management* 119: 134–142.

Zaim, O. 2004. Measuring environmental performance of state manufacturing through changes in pollution intensities: A DEA framework. *Ecological Economics* 48: 37–47.

Zhang, N., P. Zhou, and Y. Choi. 2013. Energy efficiency, CO_2 emission performance and technology gaps in fossil fuel electricity generation in Korea: A meta-frontier non-radial directional distance function analysis. *Energy Policy* 56: 653–662.

Zhao, Z. 2017. Measurement of production efficiency and environmental efficiency in China's province-level: A by-production approach. *Environmental Economics and Policy Studies* 19: 735–759.

Zhou, P., B. Ang, and K. Poh. 2006. Slacks-based efficiency measures for modeling environmental performance. *Ecological Economics* 60: 111–118.

Zhou, P., B.W. Ang, and K.L. Poh. 2008a. Measuring environmental performance under different environmental DEA technologies. *Energy Economics* 30: 1–14.

Zhou, P., B.W. Ang, and K.L. Poh. 2008b. A survey of data envelopment analysis in energy and environmental studies. *European Journal of Operational Research* 189: 1–18.

Zhou, P., B.W. Ang, and H. Wang. 2012. Energy and CO_2 emission performance in electricity generation: A non-radial directional distance function approach. *European Journal of Operational Research* 221: 625–635.

Zhou, H., Y. Yang, Y. Chen, and J. Zhu. 2018. Data envelopment analysis application in sustainability: The origins, development and future directions. *European Journal of Operational Research* 264: 1–16.

Zhu, J., and W.D. Cook. 2007. *Modeling data irregularities and structural complexities in data envelopment analysis*. New York: Springer Science + Business Media, LLC.

Zotter, K.A. 2004. "End-of-pipe" versus "process-integrated" water conservation solutions: A comparison of planning, implementation and operating phases. *Journal of Cleaner Production* 12: 685–695.

An Overview of Issues in Measuring the Performance of National Economies

Anthony Glass, Karligash Kenjegalieva, Robin C. Sickles and Thomas Weyman-Jones

1 Introduction

In this chapter, we will review the ways that economists measure the aggregate economic performance of national economies. This is the lead-in to a number of separate chapters that develop particular themes so that this chapter is intended to give an overview and anticipation of general issues that may be met in more detail subsequently. Efficiency and productivity analysis using the methodologies of data envelopment analysis and stochastic frontier analysis has made a significant contribution to this challenge after the initial research which arose in the context of the analysis of economic growth. That initial research led to the idea of measuring total

A. Glass · K. Kenjegalieva (✉) · T. Weyman-Jones
School of Business and Economics, Loughborough University, Loughborough, Leicestershire, UK
e-mail: K.A.Kenjegalieva@lboro.ac.uk

A. Glass
e-mail: A.J.Glass@lboro.ac.uk

T. Weyman-Jones
e-mail: t.g.weyman-jones@lboro.ac.uk

R. C. Sickles
Economics Department, Rice University, Houston, TX, USA
e-mail: rsickles@rice.edu

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_17

factor productivity change, TFP,[1] and its identification with an unobserved data residual representing technological progress. The contribution of efficiency and productivity analysis has been to expand our understanding of what TFP could consist of and what could drive it and how we can extend our understanding of it beyond the idea of an unexplained data residual. Amongst the critical questions in this search is the exact definition of what measure of economic performance economists should use. The conventional answer is to measure economic performance by real gross domestic product, GDP, i.e. the gross value-added measure of GDP. However, it has been frequently suggested that a broader measure of economic welfare should be used and research in this area is particularly lively now in the early part of the twenty-first century.

It is important also to be clear about what this chapter cannot do and does not do. We cannot properly survey the existing literature on the performance of national economies since there are already tens of thousands of papers on this topic and no selection could possibly give a balanced or even-handed guide to this vast literature. Nor do we plan to survey the methodologies involved in measuring the performance of national economies using efficiency and productivity analysis: by these, we mean growth accounting and the construction of index numbers, data envelopment analysis DEA, including free disposal hull methods FDH, stochastic frontier analysis SFA, stochastic non-parametric envelopment of data methods StoNED and other non-parametric regression methods. The reason is clear: the remainder of this book treats these methods in detail, and it would be foolish to offer any duplication of this material. The purpose of this chapter, once these two impractical directions are excluded, therefore arises from the title and the idea of introducing the comparative ideas and themes that a researcher into the performance of national economies might want to consider when evaluating this massive research challenge. We have deliberately organised our range of topics very widely in order to meet the challenge set for us and while we are not so naïve as to imagine that our selection will meet with wide approval, we hope that we will stimulate researchers to think broadly about the sort of issues that a non-specialist might ask about when considering the wide-ranging topic of the performance of national economies.

The most important theme that we wish to emphasise is that we have interpreted the concept of the relative *performance* of national economies

[1]We adopt the usage of representing *total factor productivity change* by the symbols TFP since that is the convention adopted elsewhere in the Handbook. Much of the macroeconomic literature simply calls this total factor productivity even though it is measured by the difference between two weighted rates of change over time. Where the reference is to the level of total factor productivity, we point this out.

very widely. Anyone familiar with the upsurge in questions by economics students about the relevance of their studies will know that neoclassical economics and the overriding dominance of GDP as the only measure of performance is under serious debate. A good example of this is the CORE project[2] which is an innovative approach to widening the economics curriculum in response to student-centred requests and which is now being widely adopted in Europe and the USA. Consequently, we devote space to examining a wide range of different concepts of economic performance including but certainly not limited to the value-added definition of GDP. In part, this reflects the expanding interest in behavioural economics, see Thaler (2018) who has emphasised how individual decision makers in reality appear to use concepts of altruism, fairness, subjective adjustment of objective frequencies, and heuristics that lead to behaviour that is at odds with the idea of maximising productivity growth.

This chapter is structured to fall broadly into two parts. The first part, which consists of Sects. 2 and 3, begins with the conventional neoclassical[3] definition of TFP using the change in the gross value-added measure of GDP in order to show what needs to be assumed to arrive at the identification of TFP with the unobserved residual that represents technological progress. There is a wide debate on why this measure of technological progress appears to have slowed down considerably in developed economies in recent years. The phrase "productivity slowdown" or "productivity gap" has become common in public discussion. Theories range from the idea that humanity has run out of new technological ideas all the way to secular stagnation meaning there is nothing worthwhile left in which to invest. We do not survey all of these ideas but we do discuss the methodological context in which they are considered.

It may of course simply be that the TFP measure used for this debate is severely at fault. This raises two types of question. First, do the conventional measures of GDP, i.e. aggregate expenditure on final goods and services or aggregate gross value-added, exclude important components of GDP? We briefly review some of the most important recent contributions to this question. Second, is GDP the appropriate indicator of national economic performance? Therefore, we follow this by a discussion of other measuring metrics

[2]See https://www.core-econ.org.

[3]The proponents of the conventional treatment of TFP as a residual use the term *neoclassical* to describe it.

of economic performance, for example, economic welfare which can present a totally different picture.

However, since the majority of the empirical work until now has focused on the measurement of economic performance in terms of GDP or GDP per capita, we treat this approach in the second part of the chapter, which consists of Sects. 4–8. We begin this part with the approach of efficiency and productivity analysis which explicitly relaxes the strong assumptions made to achieve the conventional residual TFP measure. By efficiency and productivity analysis, we mean the whole range of methodologies which flowed from the pioneering work of Farrell (1957), Charnes et al. (1978) and Aigner et al. (1977), i.e. data envelopment analysis, stochastic frontier analysis and all the subsequent developments that measure the distance of economies from their technological frontier. We show how stochastic frontier analysis and data envelopment analysis modelling has been able through the idea of TFP decomposition and the measurement of inefficiency to tell us much more about TFP than the conventional approach. In particular, we show how these methodologies are able to relax the assumptions needed for the conventional neoclassical approach and we discuss how they attempt to model the components of TFP. We can conveniently classify these methodologies into regression-based approaches[4] and programming-based approaches depending on the importance attached to errors of measurement, sampling and specification. We discover that in the regression-based approach the critical ingredient in the TFP decomposition is the computation of the elasticities of the production, input distance, output distance or cost functions. In the programming-based approach, the critical ingredients are the estimated efficiency scores under different constraints. Naturally, we leave to the other chapters in the handbook the technical details of implementing these methodologies.

We review a number of critical issues such as whether the size of national economies matters, whether there is an important role for exogenous variables in explaining the unobserved TFP residual, and the role of incentives to be efficient arising from the market structure of the economy. We will explain how programming-based approaches and regression-based approaches can model these issues and the difficulties and problems in doing so. There is an important distinction between the two broad approaches. In a regression-based methodology, the key to developing a deeper

[4]In the regression-based methodology, we concentrate on the frequentist approach and do not include discussion of Bayesian methods.

understanding of the components of TFP is by addition of further explanatory variables. In the mathematical programming methodology, the key is to develop additional constraints on the optimisation problem which is at the heart of computing the efficiency scores. Since the programming-based approach can be expressed in either primal envelopment form or dual multiplier form, adding (row) constraints to the primal involves adding (column) variables to the dual.

Following these methodological sections, the chapter turns to empirical issues, and for this we deliberately use the context of the regression-based approach, because the discussion of critical issues is, in our view, more transparent in this context than if we were to use the programming-based approach. We emphasise however that the concept of output or outputs used can include any of the performance metrics raised earlier, not only GDP. The first empirical problem we consider arises from the fact that interest in the performance of national economies is inseparable from the comparison of different national performances and this requires us to address the problem of latent heterogeneity in cross-country samples, i.e. *differences across countries.* In discussing these, we review the issue of whether the performance of national economies converges over time, or whether, as suggested by endogenous growth models, the individual performance of different countries is endogenous to the country itself.

We identify a second empirical problem in the way that technological change is modelled in efficiency and productivity analysis. The majority of studies in both the regression-based approach and the programming-based approach treat technological change as a shift over time in the complete technology frontier. This shift may be Hicks-neutral or input increasing but it assumes that all production techniques benefit simultaneously from technological change. However, there is an important literature which has a long history emphasising the idea of localised technical change in which innovation and progress applies to one or two production techniques but does not shift the whole frontier. We show that there are modelling problems for efficiency and productivity analysis in this idea but that programming-based methodology or other non-parametric approaches may offer a more fruitful starting point than conventional regression-based analysis. Finally, a third estimation issue that we identify refers to similarities amongst neighbouring countries rather than the differences between them that were discussed previously. This compels us to incorporate developments in spatial analysis into our review of the performance of national economies, and we do this in a particular example of the technological spillovers amongst neighbouring countries at the level of the aggregate production function. The issue here is

how to meld together the spatial models with the standard error term specifications, and we review some very recent contributions to this problem.

This topic of the performance of national economies is very broad indeed. We can see that there is an implicit dilemma in the topic: is it macroeconomics or is it microeconomics? Certainly, in terms of plain numbers, the volume of macroeconomics treatments of national economic performance outstrips the volume of microeconomics treatments, but efficiency and productivity analysis is essentially embedded in microeconomics. There are different ways of addressing this dilemma, but we should explain ours clearly since it will not be in agreement with some approaches that other researchers may favour. We emphasise microeconomic developments particularly in regression-based and programming-based methodologies. However, we cannot pretend that the vast macroeconomics literature on the performance of national economies does not exist or is not relevant. Therefore, we set the scene by first reviewing the key ideas from the macroeconomic literature on national economic performance so that we bring out the four critical assumptions that underlie the conventional neoclassical measures of performance: *allocative efficiency, constant returns to scale, no exogenous variable shifts and no inefficiency of performance*. This enables us to motivate the microeconomic approach embedded in efficiency and productivity analysis because each of these assumptions is relaxed by the microeconomic approach to measuring national economic performance.

Our purpose in this chapter is not to present a detailed literature survey of the vast amount of research papers on the performance of national economies[5]—that would be an impossible task today. Instead, we wish to present an analytical overview of how efficiency and productivity analysis can provide the appropriate tools for assessing national economic performance. This will therefore be an introduction to the more detailed range of topics developing this issue in the following chapters.

[5]One of the co-authors has already written a detailed survey of different stochastic frontier analysis models and specifications, Sickles et al. (2017).

2 TFP: Unobserved Data Residual Representing Technological Progress

To most macroeconomists, TFP simply means the unobserved residual in aggregate data on the gross value-added measure of GDP when account is taken of the payments to inputs or factors of production. This is identified with technological progress, the key factor in raising per capita living standards over time. This measure which is known as growth accounting has been standard since the classic papers of Solow (1957) which stated that most of the growth in per capita GDP in the USA over the first half of the twentieth century was not due to factor accumulation but was due instead to the unobserved residual which he named as technological progress, and Jorgenson and Griliches (1967) which demurred from this conclusion. In the macroeconomics literature, this measure is arrived at by the following calculation,[6] see, for example, Goodridge et al. (2016). Suppose that for sector or industry $j = 1, \ldots, J$, aggregate labour used and aggregate capital used, L_j and K_j, produce gross value added, V_j. Then the relative change in aggregate real value added is

$$\Delta \ln V = \sum_{j=1}^{J} w_j \Delta \ln V_j = \sum_{j=1}^{J} w_j v_j \Delta \ln L_j + \sum_{j=1}^{J} w_j v_j \Delta \ln K_j + \sum_{j=1}^{J} w_j \Delta \ln \text{TFP}_j \quad (1)$$

The weights v_j and w_j are respectively nominal value added in industry $j = 1, \ldots J$ as a share of aggregate value added and shares of factor cost in nominal industry value added. TFP is $\sum_{j=1}^{J} w_j \Delta \ln \text{TFP}_j$, i.e. the data residual required to ensure that the right-hand side aggregates sum to the left-hand side. This is what macroeconomists identify as technological progress. In ten Raa and Mohnen (2002), there is suggested a neat way of overcoming the problem in this growth accounting literature of using input prices as exogenous components in the weights to measure TFP when the input prices are themselves endogenous to the performance of the economy itself. ten Raa and Mohnen (2002) use shadow prices measured from an optimising model of national economic performance in which the objective is to maximise the level of final demand given the endowments and technology of the economy represented by its input-output social accounting matrix.

[6] The chapter by Fox and Diewert elsewhere in this volume addresses this issue in much more detail.

From these data, national economic performance is often defined in terms of output per worker or output per hour worked: $\Delta \ln (V/N)$ or $\Delta \ln (V/H)$ using measures of the workforce, N, or hours worked, H. The apparent downwards trend in the major developed economies, USA, Japan and the European economies including the UK, in recent years is what constitutes the productivity slowdown. There are two types of explanation of national economic performance using this approach. The first is a careful deconstruction and refinement of the labour and capital data to ensure the minimum role for TFP. The second is a range of speculations on the variability of TFP measured in this way.

However, it is important to understand that very strong implicit assumptions about the structure of the aggregate economy are needed in order to use the growth accounting approach outlined above. These include the assumptions:

- that inputs are paid the value of their marginal products and output is priced at the marginal benefit of consumption, i.e. that there is *allocative efficiency* in all markets;
- that there are *constant returns to scale* in every industry;
- that no producers display inefficiency of performance due, for example, to agency problems or behavioural patterns different from those of rational economic agents, i.e. *every producing unit is on its production frontier;*
- that *ceteris paribus* prevails, i.e. there are *no important exogenous variable changes* or changes in the market structure or regulations of the economy under study.

Goodridge et al. (2016) are careful to comment that they estimate TFP as a residual, but, they ask, what drives TFP? In theory, they say, it is technical progress, but it could also be: "*increasing returns to scale, omitted inputs, factor utilisation and cyclical effects, measurement error and a host of other factors*". In other words, the neoclassical approach to measuring TFP assumes by necessity that all of these factors are absent. As we shall see, the approach of efficiency and productivity analysis is to focus on these other factors. In this way, the efficiency and productivity analysis methods reviewed here offer a much more flexible and open way of testing large theories of the nature of economic performance. There is no shortage of such theories, e.g. the encyclopaedic summary of growth under good and bad capitalism outlined by Baumol et al. (2007). These authors offer, like others, a wide range of suggestions for enhancing economic growth which are testable using the methods of efficiency and productivity analysis but which are difficult to assess

when the neoclassical assumptions of the growth accounting approach are used. Therefore, relaxing these assumptions becomes the key to understanding how national economic performance can be compared.

Before we do this, we briefly examine the two other explanations: deconstruction of the input data and speculation about the socio-economic determinants of TFP treated as technological progress alone.

A widely cited example of the input data deconstruction approach is Gordon (2003). Suppose that we examine the measure of the rate of change of real gross value added, GDP, $\Delta \ln(V)$. We might believe that a useful decomposition is[7]

$$V = (V/H) \times (H/E) \times (E/N) \times (N/POP) \times POP \tag{2}$$

Here

V/H is gross value-added per hour worked in the sector under study;
H/E is aggregate hours worked per employee;
E/N is the employment rate—current employees as a share of the labour force;
N/POP is the labour force participation rate—those in the labour force as a proportion of the relevant population.

Only the population is regarded as a non-cyclical variable, the other ratios may all be cyclical. In Gordon (2003), the underlying trends in these ratios are identified using Hodrick-Prescott and Kalman filter time-series methods which then permit the development of socio-economic analyses of why the trends may be pointing in a particular direction. There are multiple versions of these speculative analyses in the literature. For example, with reference to the USA and other advanced economies, Baker et al. (2005) highlight demographic and population issues suggesting that populations are ageing and there are limited further reservoirs of female participation in employment because feminism is in a mature stage. To the extent that the productivity slowdown or weaker national performance is technical progress, Gordon (2016) is amongst the most prominent advocates of the argument that it has slowed because the modern age has run out of ideas. There has been a temporary boost to economic performance from the ICT-based

[7]Using an expanded identity as an analytical starting point is a popular technique for developing a new direction in research, but sooner or later it has to be supported by empirical evidence for testing theories about human behaviour.

digital revolution including the smartphone but this is ending and Gordon makes the bold claim that these innovations of the twenty-first century are as nothing compared with the great inventions of the previous 150 years: steam power, railways, natural gas pipelines, the internal combustion engine, electrical power generation and the jet engine. Why is technological progress slowing down? Gordon's explanation is that advanced economies are running into what he terms "headwinds". These include demographics associated with the retirement of ageing baby boomers leading to lower labour force participation rates. Additionally, there is an education headwind because, he argues, there is no further room for greater high-school, i.e. secondary education, completion rates. Gordon adds that inequality is worsening as the top 1% stretch away from the rest and that this reduces incentives to raise productivity generally. There are echoes of these arguments in the revival of early Keynesian ideas about secular stagnation, and related research on the long-term trend towards a falling real return on capital and consequent disincentives to invest, e.g. Lukasz and Smith (2015) who characterise the global economy as experiencing higher saving rates due to ageing populations and growing inequality, and lower returns due to falling public investment.

We might expect that there should be an important role for the shift to the digital-knowledge economy in this type of analysis, and one approach focused largely on this is the shift towards "capitalism without capital" suggested by Haskel and Westlake (2017). The starting point is their observation that investment in tangible fixed assets is becoming much less important in developed economies than what they refer to as "intangible investment" which comprises investment in design, branding, R&D, data and software. They quote the example of Microsoft which in 2006 had recorded assets that amounted to about 30% of its then market value, but 85% of these assets consisted of cash while conventional plant and equipment accounted for only 3% of the assets and 1% of the market value. They cite Microsoft as one of the first examples of capitalism without capital. Haskel and Westlake (2017) use as a critical indicator the ratio of the value of the tangible assets on a firm's balance sheet to the market value of a firm. They show that for the world's five most valuable companies, this ratio is currently (2017) below 5%; they comment that although these include the global "tech" companies this phenomenon is spreading to every sector. They argue that this makes the modern intangible-rich economy fundamentally different from one based on tangibles. Several problems arise for the measurement of performance as a result of this development. Investment in intangibles is difficult to measure in national statistics and often R&D

is simply recorded as a cost rather than a form of investment. Haskel and Westlake argue that intangible investment such as a brand, or an algorithm can be scaled up much more easily than tangible investment through the transfer of software. In addition, intangible investment has spillovers making it more difficult to stay ahead of the competition but also driving a wedge between the private and social rates of return on this form of investment. Issues such as these suggest that the way that national economic performance can be measured is likely to change radically in the future compared with the way it has been carried out up to now.

There are therefore numerous analyses in the literature that allow economists to speculate in general socio-economic terms about perceived facts of modern society, but all of them suffer from a departure from well-formulated empirical analysis and the imposition of strong assumptions about markets and behaviour and that is a gap that efficiency and productivity analysis research tries to fill. Before considering the efficiency and productivity analysis in more detail however, we must first ask whether the gross value-added measure of GDP is adequate for addressing national economic performance.

3 Is GDP the Right Way to Measure National Performance?

There are two questions to ask in this context.

Is GDP measured properly? GDP in the national accounts is gross value-added and it equals not only spending on final goods and services but also factor incomes.

Is GDP the appropriate output variable to measure?

There is considerable literature on each of these issues for which we provide a brief introduction.

There is a widely perceived idea that measured GDP excludes many important areas of economic activity, particularly in relation to mispriced goods and services. The proper definition of GDP has been a subject of debate since the development of national accounts, which were an outcome of the problems that Keynes and his followers in the USA and the UK encountered in trying to measure the level of economic activity before and during the Second World War—for a lively account of the early Keynesian efforts to define and understand GDP, see Skidelsky (2003). A key issue is the definition of the production boundary, Coyle (2014, 2017).

The production boundary[8] separates "paid-for activities in the market economy from unpaid activities" so that firms and government are considered productive but households are not (Coyle 2017). As a consequence, much of the work done largely by women in the home is not generally included in GDP—see Folbre and Nelson (2000), On the other hand, it is plain that the options for female participation in the labour force differ widely because childcare provisions vary so markedly across even the developed economies in the OECD and EU, as shown by Bettio and Plantenga (2004). Most leisure activity is excluded as well, and the digital economy is said to be having a massive but unmeasured effect on economic activity, Varian (2016). Once we start to unravel the definition of GDP, the problems of using it to measure the economic performance of national economies seem to multiply exponentially. For example, there are massive policy changes under debate and in progress to combat climate change. The achievement of a viable carbon-neutral economy is the objective of many in the environmental movement, but without a clear consensus on the social cost of carbon, we have no way of measuring the benefits in GDP terms of the success or otherwise of environmental policy.

Consequently, economists have for decades argued that GDP is an inadequate measure of the economic performance of nations and have sought to develop alternative measures of national economic welfare. There have been many suggestions for a welfare or even a "happiness"-based index instead, see Helliwell et al. (2012). Particularly important have been suggestions by international bodies like the UN which has developed its own human development index, HDI that includes measures of education and health. Many of the suggested substitute measures such as "happiness" are based on survey responses, and Helliwell et al. (2012) are the most widely cited of these. The initial observation that commenced this line of research is the Easterlin (1974) paradox that states that at any point in time richer people appear to be happier than poorer people but over time society does not appear to become happier as it becomes richer. Easterlin's explanation is that individuals use relative income levels to evaluate their well-being but if these stay constant over time happiness is unchanged. The contention of Helliwell, Layard and Sachs is that happiness differs over time and across societies for

[8]The idea of the production boundary separating productive and non-productive services goes back to Adam Smith (1776) where Smith famously distinguished the output of productive labour from that of non-productive labour whose "services generally perish in the very instant of their performance, and seldom leave any trace or value behind". Many professions fell into this category, according to Smith, including the menial servant, the Sovereign, men of letters of all kinds, buffoons and opera singers.

identifiable reasons, and it may be alterable by public policy. Their 2012 report used the Gallup World Poll, the World Values Survey, the European Values Survey and the European Social Survey from 2005 to 2011 to compile a broad happiness index. For example, the Gallup World Poll asked 1000 people aged 15 or over in 150 countries to evaluate the quality of their lives on an ascending score from 0 to 10. For the world as a whole weighted by population, the modal category, i.e. the category with the largest number of people reporting (26.2%), was 5, exactly the mid-point. In categories 6–10, there was a further 42.9% of respondents so that 69.1% of the total reported that they were not below the mid-point of the happiness scale. For North America, Australia and New Zealand, the modal life satisfaction category was 8 with 92.9% of respondents reporting that their life satisfaction was not below the mid-point (5). By contrast in sub-Saharan Africa, only 47.4% of respondents were not below mid-point category 5. Clearly level of development with all of its associated implications plays a major role in the relative evaluation of happiness. In their analysis of responses, Helliwell Layard and Sachs identified key categories affecting life evaluations as: *work* (employment and quality); *social capital* (trust, freedom, equality); *values* (altruism, materialism, environment); *health* (mental, physical); *family*; *education level*; *gender.* For example, improvements in the nature of work or the support for social capital and health were evaluated as being worth several multiples of a 30% increase in income. Based on results like these, Helliwell Layard and Sachs noted that changes in these factors can be brought about by policy reform, offering considerable scope for rich and deep analysis on the relative performance of national economies.

However, there has also been a consistent strand of economic research that attempts to measure economic welfare amongst nations empirically rather than subjectively. In Jones and Klenow (2016), for example, there is a detailed empirical study that compares the performance of a wide range of countries on a measure of economic welfare determined by an equivalent consumption metric. Their aim is to determine how an easily computable measure of economic welfare correlates with GDP as a measure of economic performance. It is interesting to examine this example of much of the recent work on the usefulness of GDP as a measure of national economic performance. Jones and Klenow imagine an individual living in an arbitrarily chosen country and drawing his/her life experiences from that country's distributions of *consumption, work-leisure trade-off opportunities, inequality and life expectancy*. Using simple logarithmic assumptions about preferences, they construct from observed macro- and micro-data a measure of utility for that individual in that country. They then construct a variable: the "consumption

equivalent measure of standard of living" which is the factor *lambda* which if applied to the random draws of consumption, leisure and life expectancy from the distributions applying in the USA would make that individual indifferent between living in the USA and his/her original country. The factor *lambda* is the number which multiplicatively reduces the level of consumption of a US citizen sufficiently to provide a level of utility equivalent to the citizen of another country when utility depends on consumption, inequality and leisure, and when consumption in each country is a randomly distributed variable with a mean and variance particular to the country in question. In other words, the proportion of USA consumption—given the leisure, mortality and inequality in the USA—which would provide the same expected utility as the values elsewhere.

Jones and Klenow provide a simple example.[9] They postulate an intercept level of utility, $\bar{u}$, e.g. the lifetime subsistence level of consumption or value of life, and concentrate on two key variables: the first is consumption per capita, C, which is the individual's random draw from the consumption distribution for the country in which he/she lives, and the second is the utility of leisure time, $v(l)$, drawn from the leisure distribution in the country. The flow of utility is

$$u(C, l) = \bar{u} + \log C + v(l) \tag{3}$$

They assume that consumption is log-normally distributed, a result often found to describe all but the top percentile of the income distribution in many countries:

$$\log C_i \sim N\left(\mu_i, \sigma_i^2\right) \tag{4}$$

where $\mu = E(\log C)$ and $\sigma^2 = \text{var}(\log C)$.

Jones and Klenow parameterise the mean of consumption in country i as $E(C_i) = c_i$, then, using the properties of the log-normal distribution, they are able to write[10]:

$$E(C) = \exp\left(\mu + \frac{1}{2}\sigma^2\right) = c \Rightarrow \log E(C) = \mu + \frac{1}{2}\sigma^2 = \log c \tag{5}$$

[9]Jones and Klenow present a complex analysis of which this is the simplest example assuming a zero discount rate for utility of consumption and a zero growth rate for consumption.

[10]The derivation of these results is compressed in Jones and Klenow (2016) so we have expanded the explanation.

i.e. after rearranging Eq. (5) and using (4):

$$E(\log C) = \mu = \log E(C) - \frac{1}{2}\sigma^2 = \log c - \frac{1}{2}\sigma^2 \tag{6}$$

Assuming that typical life expectancy in any year for a citizen in this country i is e_i, Jones and Klenow then write, in their simplest case, that a citizen's expected lifetime utility is the product of the flow of utility multiplied by life expectancy:

$$U_i = e_i\left(\bar{u} + \log c_i + v(l_i) - \frac{1}{2}\sigma_i^2\right) \tag{7}$$

This tells us that welfare of the typical citizen in this country is increasing in life expectancy, increasing in consumption per person, increasing in the utility of leisure available per person but decreasing in the variance of consumption per person, which is a measure of the inequality of the distribution of consumption per person.

Now for the case $i = \text{USA}$ multiply c_{USA} by λ_i, the multiplier by which U_{USA} must be reduced to yield the level of welfare that is equivalent to that of a citizen living in country i.

$$U_{\text{USA}}(\lambda_i) = U_i(1) \tag{8}$$

In other words, find λ_i that satisfies

$$e_{\text{USA}}\left(\bar{u} + \log(\lambda_i c_{\text{USA}}) + v(l_{\text{USA}}) - \frac{1}{2}\sigma_{\text{USA}}^2\right) = e_i\left(\bar{u} + \log c_i + v(l_i) - \frac{1}{2}\sigma_i^2\right) \tag{9}$$

The result is

$$\begin{aligned}\log \lambda_i = {} & [(e_i - e_{\text{USA}})/(e_{\text{USA}})]\left(\bar{u} + \log c_i + v(l_i) - \frac{1}{2}\sigma_i^2\right) + \left[\log c_i - \log c_{\text{USA}}\right] \\ & + [v(l_i) - v(l_{\text{USA}})] - \frac{1}{2}\left(\sigma_i^2 - \sigma_{\text{USA}}^2\right)\end{aligned} \tag{10}$$

This "consumption equivalent measure of standard of living" therefore consists of four additive terms for each country:

- Relative life expectancy in country i compared with USA weighted by the mean flow of utility of consumption and leisure in country i
- Relative mean consumption compared with USA
- Relative utility of leisure time compared with the USA

- Relative variance of consumption compared with the USA, which is a measure of consumption inequality.

In general, $\log \lambda_i$ will be negative so that $\lambda_i < 1$ due to the dominance of the second term, the gap between the county's per capita consumption and that of the USA. However, as the other terms have an impact, for some countries $\log \lambda_i$ will approach zero so that $\lambda_i \cong 1$. In other words, while many countries will have consumption per capita much lower than in the USA some may have higher leisure time, higher life expectancy and a more equitable distribution of income all of which contribute positively to the citizen's welfare in the Jones-Klenow social welfare function. Therefore, we may expect that compared to the ranking by GDP or consumption per capita, the consumption equivalent welfare measure may show that some countries rank equally highly with the USA in terms of national welfare performance but that others may be much more worse off than the raw GDP data indicate.

Jones and Klenow draw on the research literature to parameterise these components in particular using a value for the Frisch elasticity of labour supply of 1 which implies disutility from working rises with the square of the number of hours worked so that $v(l)$ then depends on the real wage and the marginal tax rate of labour income. Constructing λ_i for a wide range of countries provides a set of important results for the evaluation of GDP as a measure of comparative national performance compared with the consumption equivalent measure of welfare.

- The correlation between GDP per capita and consumption equivalent welfare is very high, of the order of 0.95–0.98.
- In Western Europe, living standards are much closer to the USA than income per capita suggests due to longer lives with more leisure.
- In most developing countries, welfare is much lower than income per capita due mainly to shorter lives with more inequality.
- Economic growth in consumption equivalent welfare (except in sub-Saharan Africa) is 50% higher than growth in GDP per capita due to declining mortality.

Jones and Klenow are conscious of leaving out other aspects of welfare in which they include morbidity, environmental quality, crime, political freedom and intergenerational altruism; nevertheless, this example of a growing literature indicates how the measurement of the performance of national economies opens up a massive range of modelling developments. It is

possible, for example, to consider the consumption equivalent measure of standard of living as providing an alternative conception of the frontier of national economic performance to which further efficiency and productivity analysis could then be applied. In general, efficiency and productivity analysis of the "happiness frontier" or welfare frontier is a largely unexplored area.

4 National Economic Performance: Programming Analysis

From this point on we take up the second part of the chapter and focus on using real value-added GDP as the key measure of performance of national economies so that the estimation of TFP is the central preoccupation of the analysis. In this section, we concentrate on the programming approach to measuring TFP, generally known by the generic name of data envelopment analysis.

The paper by Farrell (1957) and the comments by Winsten (1957) contributed hugely to the development of efficiency and productivity analysis, and it is interesting that the initial example related to efficiency in aggregate agricultural production of the USA. However, it can be said that the pioneering paper in the application of efficiency and productivity analysis to measuring the performance of national economies is Färe et al. (1994). This paper used data envelopment analysis to evaluate productivity change across different countries and introduced two major changes to the assumptions required by the neoclassical growth accounting method. Constant returns to scale was replaced by the possibility of variable returns to scale and the assumption that every country was on the international production frontier was replaced by the possibility that countries could display inefficiency of performance. In this way, efficiency and productivity analysis moved on from the conventional neoclassical macroeconomic approach to measuring TFP. Färe, Grosskopf, Norris and Zhang used data envelopment analysis with variable returns to scale to develop Malmquist indices of TFP. Subsequently, Ray and Desli (1997) refined the analysis on how the effect of variable returns to scale should be measured. In ten Raa and Shestalova (2011), the Solow residual concept is neatly reconciled with the data envelopment analysis approach to productivity measurement by embedding it in an input-output analysis. This approach which makes uses of duality and shadow prices offers a potentially interesting way to conceptualise the theoretical measurement of TFP.

The distance function contains the same information about the technology as the production function. Consequently Caves et al. (1982) by adopting a normative approach rather than the neoclassical data residual approach show that productivity growth can be represented by a Malmquist index defined to be the ratio of the values of an output distance function after an event to the value of an output distance function before the event. The output distance function directly measures Farrell radial efficiency. The resulting index $M > 1$ if there is positive productivity growth. Färe et al (1994) developed this in several ways. First, by assuming that the producing unit need not be on the transformation surface either before or after the event. In other words, the producing unit could be technically inefficient despite the existence of productivity growth. The possibility of building in a measure of inefficiency allows the researcher to decompose the shift in the producing unit's position into two separate components: the efficiency change effect and the technical change effect. Secondly, Färe et al derived the equivalent Malmquist indices for the input distance function which measures the inverse of the Farrell radial efficiency, so that to maintain the convention that the resulting index $M > 1$ if there is positive productivity growth the inverses of the input distance functions, i.e. the Farrell radial technical efficiency scores were used. Finally, Färe et al. addressed the issue of returns to scale by defining the Malmquist index for distance function values assuming constant returns and a separate Malmquist index for distance functions assuming variable returns to scale. The difference between the two technologies is defined by the description of the technology as a convex cone in the case of constant returns and a convex hull in the case of variable returns.

We isolate two periods for comparison: t and $t + 1$, representing the before and after positions relative to a productivity change. We need to compare the value of the distance function at $t + 1$ to its value at t, but there is the option of choosing the period t or the period $t + 1$ output possibility set as the reference technology. For example, Färe et al. (1994) use the geometric mean of these two reference sets as the reference technology. We observe the inputs and outputs at each of these periods and set up the corresponding programming analysis with input-orientated radial efficiency measures (θ). Non-radial measures of efficiency can also be incorporated into developments of the distance function approach.

The output distance function is defined for a given technology such that a vector of inputs $\mathbf{x}' = (x_1 \ldots x_K)$ can make a vector of outputs $\mathbf{y}' = (y_1 \ldots y_R)$. The technology set is

$$T = \{\mathbf{x}, \mathbf{y} : \mathbf{x} \text{ can make } \mathbf{y}\} \tag{11}$$

The output distance function is the smallest positive scalar divisor δ of a bundle of the production unit's outputs $\mathbf{y}$ such that $(\mathbf{y}/\delta)$ is in the technology set, T.

$$D_O(\mathbf{x}, \mathbf{y}) = \min\{\delta : (\mathbf{x}, \mathbf{y}/\delta) \in T\} \leq 1 \tag{12}$$

A piecewise linear representation of the technology of production with constant returns to scale is.

$$T^r = \{\mathbf{x}, \mathbf{y} : \mathbf{X}\lambda \leq \mathbf{x}, \mathbf{Y}\lambda \geq \mathbf{y}\} \tag{13}$$

The matrices $\mathbf{X}, \mathbf{Y}$ represent all of the observed data in the efficiency measurement exercise and the vectors $\mathbf{x}, \mathbf{y}$ represent one particular country. The output distance function can be measured by the Farrell radial efficiency, δ, of each country's outputs:

$$D_O(\mathbf{x}, \mathbf{y}) = \min\left\{\delta : (\mathbf{x}, \mathbf{y}/\delta) \in T^r\right\} = \min\{\delta : \mathbf{X}\lambda \leq \mathbf{x}, \mathbf{Y}\lambda \geq \mathbf{y}/\delta\} \tag{14}$$

There is an equivalent approach to the input distance function defined as the largest scalar divisor ρ of a bundle of inputs $\mathbf{x}$ such that $(\mathbf{x}/\rho)$ is still in the technology set which leads to a Farrell radial efficiency measure, θ, of the inverse input distance function:

$$D_I(\mathbf{x}, \mathbf{y}) = \max\left\{\rho : (\mathbf{x}/\rho, \mathbf{y}) \in T^r\right\} = \min\{\theta : \mathbf{X}\lambda \leq \theta\mathbf{x}, \mathbf{Y}\lambda \geq \mathbf{y}\} \tag{15}$$

In the both cases of the output and input distance function, the assumption of variable rather than constant returns to scale is implemented by adding the constraint $\mathbf{e}'\boldsymbol{\lambda} = 1$ to the piecewise linear representation of the technology, where $\mathbf{e}$ is a vector of ones.

We use the notation $\theta_C^{t+1,t}$ to represent the input-orientated measure of radial efficiency with constant returns to scale for a country observed in period t relative to the technology prevailing in period $t+1$ while $\theta_V^{t+1,t}$ refers to the variable returns to scale version.

Scale efficiencies are given by the ratio of the Farrell radial efficiency under CRS to the Farrell radial efficiency under VRS. There are four measures: respectively scale efficiency for the observation in period t with reference to the period t technology, scale efficiency for the observation in period t with reference to the period $t+1$ technology, scale efficiency for the observation in period $t+1$ with reference to the period $t+1$ technology, and scale efficiency for the observation in period $t+1$ with reference to the period t technology:

$$\begin{aligned}\sigma^t\left(\mathbf{x}_0^t,\mathbf{y}_0^t\right) &= \frac{\theta_C^{t,t}}{\theta_V^{t,t}}\,\sigma^{t+1}\left(\mathbf{x}_0^t,\mathbf{y}_0^t\right) = \frac{\theta_C^{t+1,t}}{\theta_V^{t+1,t}}\,\sigma^{t+1}\left(\mathbf{x}_0^{t+1},\mathbf{y}_0^{t+1}\right)\\ &= \frac{\theta_C^{t+1,t+1}}{\theta_V^{t+1,t+1}}\,\sigma^{t}\left(\mathbf{x}_0^{t+1},\mathbf{y}_0^{t+1}\right) = \frac{\theta_C^{t,t+1}}{\theta_V^{t,t+1}}\end{aligned} \tag{16}$$

This produces a scale decomposition:

$$\text{SEC} = \left[\frac{\sigma^t\left(\mathbf{x}_0^{t+1},\mathbf{y}_0^{t+1}\right)}{\sigma^t\left(\mathbf{x}_0^t,\mathbf{y}_0^t\right)}\,\frac{\sigma^{t+1}\left(\mathbf{x}_0^{t+1},\mathbf{y}_0^{t+1}\right)}{\sigma^{t+1}\left(\mathbf{x}_0^t,\mathbf{y}_0^t\right)}\right]^{1/2} \tag{17}$$

Malmquist indices can then be defined for CRS or VRS technology

$$M_{IC}\left(x_0^t,y_0^t,x_0^{t+1},y_0^{t+1}\right) = \left[\frac{\theta_C^{t+1,t+1}}{\theta_C^{t,t}}\right]\left[\frac{\theta_C^{t,t}}{\theta_C^{t+1,t}}\,\frac{\theta_C^{t,t+1}}{\theta_C^{t+1,t+1}}\right]^{1/2} \tag{18}$$

$$M_{IV}\left(x_0^t,y_0^t,x_0^{t+1},y_0^{t+1}\right) = \left[\frac{\theta_V^{t+1,t+1}}{\theta_V^{t,t}}\right]\left[\frac{\theta_V^{t,t}}{\theta_V^{t+1,t}}\,\frac{\theta_V^{t,t+1}}{\theta_V^{t+1,t+1}}\right]^{1/2} \tag{19}$$

Each index can be decomposed into a measure of efficiency change, EFC, the first ratio in square brackets, and technical change, TEC, the second ratio in square brackets.

Finally, we have a relationship between the indices:

$$M_{IC}\left(x_0^t,y_0^t,x_0^{t+1},y_0^{t+1}\right) = M_{IV}\left(x_0^t,y_0^t,x_0^{t+1},y_0^{t+1}\right)\times\text{SEC} \tag{20}$$

This provides us with a complete decomposition into efficiency change EFC, technical change TC and scale efficiency change, SEC:

$$M = \text{EFC}\times\text{TC}\times\text{SEC} \tag{21}$$

Fare et al. (1994) and Ray and Desli (1997) applied the analysis to an international sample using data on real GDP, labour and capital inputs from the Penn World Tables, Summers and Heston (1991) and this procedure has become commonplace in international productivity comparisons using the efficiency and productivity analysis approach. In summary, this early

literature was able to introduce two ways in which the neoclassical assumptions could be relaxed by employing data envelopment analysis. The data envelopment analysis assumes that countries can be below the international production frontier and that they can operate with technologies that display variable returns to scale. The procedure for doing this is straightforward once we abandon the neoclassical approach and specify the ideas associated with the distance function and Farrell radial efficiency.

There are many ways in which this initial work on international TFP comparisons used efficiency and productivity analysis. For example, Milner and Weyman-Jones (2003) in a study confined to developing nations also drew on the Penn World Tables to measure the radial efficiency component of different countries and to relate that to different measures of country heterogeneity, thereby combining the neoclassical approach that focused on determinants of differences in national GDP performance and the efficiency and productivity analysis approach which modelled the technology as a performance measure that varied with returns to scale. In their study of developing nations, Milner and Weyman-Jones (2003) looked for possible determinants of the measured Farrell radial efficiency scores of different countries that used inputs of labour, capital and agricultural land to generate real GDP. The explanatory factors they used in a 2-stage analysis included relative country size, per capita income, education level, health level, industrialisation, degree of democracy, trade openness. There have been many further advances in the data envelopment analysis approach to international productivity comparisons. Giraleas et al. (2012) demonstrated that the data envelopment analysis approach using Malmquist indices performed particularly well in simulation studies when compared against the neoclassical growth accounting and deterministic regression-based frontiers. Since the development of the programming approach by Färe, Grosskopf, Norris and Zhang, there has been a massive expansion in the number of data envelopment analysis studies of the Malmquist estimation of TFP for a multiplicity of economies.

5 National Economic Performance: Regression-Based Analysis

In assessing the performance of national economies from the point of view of efficiency and productivity analysis, we have the choice of measuring either efficiency levels or productivity changes across space and time. The major part of the literature concentrates on productivity comparisons and

changes because simply focusing on the measured distance to a frontier does not bring out the major factors that could be important in decomposing the changes in productivity over time. In Sect. 4, we showed that by using data envelopment analysis to construct normative Malmquist indices of TFP, we are able to relax two critical assumptions of the conventional neoclassical approach: the assumption of constant returns to scale and the assumption that every country is on the international production frontier. The other two assumptions of the neoclassical approach can also be investigated. These are that allocative efficiency prevails and that *ceteris paribus* is invoked—i.e. there are no exogenous variable shifts to take into consideration. The neoclassical approach gets around the second of these two requirements by *ad hoc* qualitative speculation about long-term socio-economic trends including demographics supported by detailed deconstruction of the data on inputs and GDP used in the traditional index number approach. We have now seen how efficiency and productivity analysis in the form of data envelopment analysis can address issues that the neoclassical approach cannot. It is natural to ask whether data envelopment analysis could also contribute to the relaxation of the allocative efficiency assumption and the *ceteris paribus* assumption. It is certainly the case that a vast amount of useful data envelopment analysis has addressed these issues as well. Allocative efficiency has been researched in the data envelopment analysis approach since the original Farrell contribution and data envelopment analysis models can be developed using input price data to capture allocative inefficiency, see, e.g. Bogetoft and Otto (2011). It is also true that data envelopment analysis can be redesigned to accommodate additional shift factors representing the role of exogenous variables—this is done by adding constraint rows to the primal envelopment DEA programmes or equivalently adding variable columns to the dual DEA multiplier problems. In addition, some progress towards the inclusion of idiosyncratic error in the form of sampling error can also be made using bootstrapping approaches, Simar and Wilson (2007). However, these issues can also be addressed using stochastic frontier analysis and arguably in the context of comparing the performance of national economies rather than individual decision-making organisations, the stochastic frontier analysis approach offers clearer and more direct methods of analysis. Consequently, in continuing our discussion of the relaxation of the conventional neoclassical approach to TFP we turn to stochastic frontier analysis.

There are different ways of deriving the full TFP formulae for different representations of the technology. One procedure as we saw is to start from the generalised Malmquist index form shown in Eqs. (20) and (21) in which the Malmquist index of distance function values, which can be decomposed

into a technological shift and a frontier catch-up, is adjusted by a scale factor to take account of non-constant returns to scale. Orea (2002) showed how this can be developed in a stochastic frontier analysis framework to generate a Tornqvist index of total factor productivity change, TFP, by using the empirically estimated elasticities and the key idea of the quadratic identity lemma due to Diewert (1976). Coelli et al. (2003) apply this to the production function, input distance function and multi-product cost function representations of the technology.

Another derivation starts from the basic properties of a TFP index. An index of TFP is the weighted growth rates of outputs minus the weighted growth rates of inputs. Two of the most important properties of the weights are monotonicity and proportionality. Monotonicity requires that the weighted output growth rates and input growth rates are chosen so that higher output and lower input unambiguously improve TFP. This requires that in an empirical application based on regression analysis the elasticities must all be adjusted to have non-negative signs. Proportionality requires that the weights in the output and input growth indices add to unity. We can apply these requirements to the empirical estimation of the single output production function, the multi-product input distance function, the multi-product output distance function and the multi-product cost function to derive TFP indices from each representation of the technology. We first differentiate our functional representation with respect to time to obtain the proportional rates of change of the outputs and the inputs then we choose a functional form to estimate by regression procedures so as to generate the required elasticity weights.

We apply this as follows in Table 1. The analysis here is based on the approach of Lovell (2003) who illustrated the output distance function. For each function: production, input distance, output distance and cost, we derive the elasticity weights from the logarithmic form of the function by total differentiation with respect to time. We use these elasticities as the output and input weights ensuring that they satisfy monotonicity and proportionality. In the case of the input distance function, for example, we must take care that the output elasticities which are negative when estimated are changed in sign to ensure monotonicity and in the case of the output distance function where the input elasticities are negative the same adjustment applies. The proportionality property is ensured by adjusting the elasticities by the elasticity of scale value for the function in question, respectively E, E^I, E^O, E^C for the production, input distance, output distance and cost functions. When $E, E^I, E^O, E^C > 1$ there are increasing returns or economies of scale, when $E, E^I, E^O, E^C = 1$ there are constant returns and when

Table 1 Functional forms and decompositions of TFP for different representations of the technology

Function for estimation	Form for estimation	Form for TFP components satisfying monotonicity and proportionality
Production: Elasticity of scale: $E = \sum_k \varepsilon_{x_k}$	$\ln y = \ln f(x', \mathbf{z}', t) - u + v$ $\varepsilon_{x_k} = \partial \ln f / \partial \ln x_k \geq 0$	$\ln y = \ln f(\mathbf{x}', \mathbf{z}', t) - u, u \geq 0$ $[\dot{y} - (1/E)\boldsymbol{\varepsilon}_{\mathbf{x}}'\dot{\mathbf{x}}] = TFP = (1 - 1/E)\boldsymbol{\varepsilon}_{\mathbf{x}}'\dot{\mathbf{x}} + \boldsymbol{\varepsilon}_{\mathbf{z}}'\dot{\mathbf{z}} + \varepsilon_t - (\mathrm{d}u/\mathrm{d}t)$ Decomposition: $TFP = SEC + EXC + TC + EFC$
Input distance, $\sum_k \varepsilon_{x_k} = 1$ Elasticity of scale: $E^I = \left(-\sum_r \varepsilon_{y_r}^I\right)^{-1}$	$-\ln x_K = \ln D_I((1/x_K)\mathbf{x}', \mathbf{y}', \mathbf{z}', t) - u + v$ $\varepsilon_{x_k}^I = \partial \ln D_I / \partial \ln (x_k/x_K) \geq 0$ $\varepsilon_{y_r}^I = \partial \ln D_I / \partial \ln y_r \leq 0$	$\ln D_I(\mathbf{x}, \mathbf{y}, \mathbf{z}, t) - u = 0, u \geq 0$ $\left[-E^I \boldsymbol{\varepsilon}_{\mathbf{y}}^{I\prime}\dot{\mathbf{y}} - \boldsymbol{\varepsilon}_{\mathbf{x}}^{I\prime}\dot{\mathbf{x}}\right] = TFP = (1 - E^I)\boldsymbol{\varepsilon}_{\mathbf{y}}^{I\prime}\dot{\mathbf{y}} + \boldsymbol{\varepsilon}_{\mathbf{z}}^{I\prime}\dot{\mathbf{z}} + \varepsilon_t^I - \mathrm{d}u/\mathrm{d}t$ Decomposition: $TFP = SEC + EXC + TC + EFC$
Output distance, $\sum_r \varepsilon_{y_r} = 1$ Elasticity of scale: $E^O = \left(-\sum_k \varepsilon_{x_k}^O\right)$	$-\ln y_R = \ln D_O(\mathbf{x}', (1/y_R)\mathbf{y}', \mathbf{z}', t) + u + v$ $\varepsilon_{y_r}^O = \partial \ln D_O / \partial \ln (y_r/y_R) \geq 0$ $\varepsilon_{x_k}^O = \partial \ln D_I / \partial \ln x_k \leq 0$	$\ln D_O(\mathbf{x}, \mathbf{y}, \mathbf{z}, t) + u = 0, u \geq 0$ $\left[\boldsymbol{\varepsilon}_{\mathbf{y}}^{O\prime}\dot{\mathbf{y}} + \left(1/E^O\right)\boldsymbol{\varepsilon}_{\mathbf{x}}^{O\prime}\dot{\mathbf{x}}\right] = TFP = \left(1/E^O - 1\right)\boldsymbol{\varepsilon}_{\mathbf{x}}^{O\prime}\dot{\mathbf{x}} - \boldsymbol{\varepsilon}_{\mathbf{z}}^{O\prime}\dot{\mathbf{z}} - \varepsilon_t^O - \mathrm{d}u/\mathrm{d}t$ Decomposition: $TFP = SEC + EXC + TC + EFC$
Dual expenditure (cost) for industry (value-added sector) $j = 1, \ldots, J$, $\sum_k \eta_{w_k} = 1$ Elasticity of scale: $E^C = (\sum_r \eta_{y_r})^{-1}$	$\ln (C_j/w_K) = \ln c^j((1/w_K)\mathbf{w}', \mathbf{y}', \mathbf{z}', t) + u + v$ $\eta_{w_k} = \partial \ln c^j / \partial \ln (w_k/w_K) \geq 0$ $\eta_{y_r} = \partial \ln c^j / \partial \ln y_r \geq 0$	$\ln (\mathbf{w}'\mathbf{x})_j = \ln c^j(\mathbf{w}', \mathbf{y}', \mathbf{z}', t) + u, u \geq 0$ $\left[E^C \boldsymbol{\eta}_{\mathbf{y}}'\dot{\mathbf{y}} - \mathbf{s}'\dot{\mathbf{x}}\right] = TFP = \left(E^C - 1\right)\boldsymbol{\eta}_{\mathbf{y}}'\dot{\mathbf{y}} + (\mathbf{s} - \boldsymbol{\eta}_{\mathbf{w}})'\dot{\mathbf{w}} - \boldsymbol{\eta}_{\mathbf{z}}'\dot{\mathbf{z}} - \eta_t - (\mathrm{d}u/\mathrm{d}t)$ $\mathbf{s}_k$: observed cost share, $\sum_k s_k = 1$ Decomposition: $TFP^j = SEC + AEC + EXC + TC + EFC$ Aggregate: $TFP = \sum_j v_j TFP^j$

$E, E^I, E^O, E^C < 1$ there are decreasing returns or diseconomies of scale. These elasticity of scale formulae are derived for the distance functions in Fare and Primont (2012) and for the cost function in Panzar and Willig (1977). The resulting TFP index will therefore be decomposable into four components: scale efficiency change, SEC, the change due to exogenous variables, EXC, the technological change or frontier shift effect, TC, and the efficiency change component, EFC, derived from the inefficiency component of the error term in the empirical estimation form.

To construct the TFP measures shown in Table 1, we have identified for each function an elasticity weighted average of rates of output growth minus an elasticity weighted average of rates of input growth. Monotonicity requires that the TFP measure increases if outputs increase and decreases if inputs increase. This requirement determines the sign properties of the components. Proportionality requires that the different elasticity weights applying to outputs and to inputs each sum to one. This requirement is partly satisfied when the functional form representing the underlying technology has some form of linear homogeneity property; otherwise, it must be satisfied by making a scale adjustment to the elasticity weighted rates of change of outputs or inputs. In stochastic frontier analysis, we find that different decompositions are available depending on the way in which we choose to model the technology and the behaviour of producers.

The simplest place to begin is the aggregate production function relating our preferred output measure, y, e.g. the real gross value-added estimate of GDP to the economy's aggregated inputs: $x_1, \ldots, x_K$, the exogenous variables held constant under the *ceteris paribus* assumption: $z_1, \ldots, z_L$ and time representing the possibility of technological progress, t. Write the stochastic aggregate production function as

$$y = f(x_1, \ldots, x_K, z_1, \ldots, z_L, t)\exp(v - u) \tag{22}$$

The error term has as usual two components: v is a two-sided symmetrically distributed zero-mean random variable representing idiosyncratic error which is usually assumed to encompass all the measurement error, sampling error and specification error in the data generating process; u is an asymmetrically distributed non-negative random variable with its distribution truncated at zero so that it has a positive expected value which is assumed to encompass the inefficiency of producer performance. In this way, we arrive at a measure of stochastic efficiency of performance

$$0 < TE = \mathrm{e}^{-u} = y/f(x_1, \ldots, x_K, z_1, \ldots, z_L, t)\exp(v) \leq 1 \tag{23}$$

To arrive at a productivity change measure, we must take the logarithmic derivative of the initial aggregate production function with respect to time. Write

$$\dot{y} \equiv \partial \ln y/\partial t = (\partial y/\partial t)/y \tag{24}$$

We use the same convention for all of the other variables.

In Table 1, we show in the first row the TFP decomposition for the aggregate production function representation of the technology. The term $E = \sum_k \varepsilon_{x_k}$ is the elasticity of scale (<1, =1, >1, according as there are decreasing, constant and increasing returns to scale). The left-hand side of the TFP expression (in square brackets) is the growth rate of output minus the weighted growth rates of the inputs with the weights designed to sum to 1. Therefore, it is by definition a measure of total factor productivity change, TFP. The terms on the right-hand side represent first the growth rate of inputs adjusted by the elasticity of scale, which is scale efficiency change, SEC, second the weighted growth rates of the exogenous variables, EXC, third the growth of output over time when all other variables are held constant, i.e. technological progress or technical change, TC, and fourth the rate of decline of inefficiency over time, i.e. efficiency change, EFC. Each of these terms depends on knowledge of the relevant production function elasticities.

Consequently, the stochastic frontier analysis has permitted a generalisation of the previous expressions for TFP to give us:

$$\text{TFP} = \text{SEC} + \text{EXC} + \text{TC} + \text{EFC} \tag{25}$$

In this way, the stochastic frontier analysis has relaxed three of the four key assumptions of the conventional neoclassical growth accounting approach—we no longer assume constant returns to scale or that exogenous factors must be held constant, or that all producers are on their respective production frontiers. These factors can be added to technological progress as components of TFP so long as we are able to estimate them from the available data. We do this by econometric estimation of the functional form shown in the second column of Table 1, and we choose a functional form from which it is possible to extract the relevant elasticity information.

We can derive a productivity decomposition TFP for each of our functional forms using the log-derivatives as we did above, and the second, third and fourth rows of Table 1 show the corresponding TFP decompositions.[11]

[11]Key references on these ideas are Bauer (1990), Orea (2002) and Coelli et al. (2003).

Note that in each case, the left-hand side of the TFP expression contains the definition of TFP and it is required to be the difference between a weighted sum of the log-output changes minus a weighted sum of the log-input changes with the weights summing to 1 in each case. In the case of the input distance or output distance, this requires the use of a homogeneity property and a rescaling by the corresponding measure of the elasticity of scale. For the input distance, the input elasticities must sum to 1 by the homogeneity property, and it is the output elasticities which are rescaled on the left-hand side. In the output distance, the output elasticities sum to 1 by homogeneity and the input elasticities are rescaled on the left-hand side.

None of these decompositions contains a measure of allocative inefficiency. To achieve this, we need to introduce input prices: $w_1, \ldots, w_K$ and therefore move to a dual expenditure function or cost function for each industry or sector of gross value added separately. In the case of the dual expenditure function, the cost-output elasticities on the left-hand side of the TFP decomposition are rescaled and the cost-input elasticities are weighted by their actual cost shares, s_k which must sum to 1.

Table 1 is completed by showing the TFP decomposition in this cost function case for industry j and this may be aggregated into an overall TFP decomposition using the industry weights as indicated in the conventional growth accounting approach. It is from this last row of Table 1 that we are able to incorporate an allowance for allocative efficiency change, AEC, which is measured by the log-input-price changes weighted by the difference between the actual cost shares and the optimal cost minimising cost shares, see Bauer (1990), and Orea (2002).

The procedure for estimating the TFP decomposition then proceeds as follows. Since this form of research is often used to compare different national economic performances across countries and across time, we will illustrate with panel data.

- Select a function to represent the technology and the chosen TFP decomposition.
- Select a functional form for the function.
- Estimate the functional form using stochastic frontier analysis so that the efficiency change component, EFC is included. In the case of the input and output distance functions, the homogeneity property must be imposed to make estimation feasible by identifying the dependent variable. In the production function case, homogeneity is not assumed unless constant returns to scale are imposed. In the cost function homogeneity in input prices can be imposed or it may simply be tested on the

estimated equation without homogeneity. In each case, the relevant estimated elasticities from the homogeneous and non-homogeneous form are the same.
- The left-hand side of the TFP decomposition defines the measure of TFP and is not to be calculated, since it is by definition equal to the right-hand side. Instead, the elasticity estimates are used to calculate each of the right-hand side components.
- The elasticity estimates are used to multiply the log-variable changes to arrive at the corresponding TFP decomposition. However, all of the analysis so far has assumed continuous functions and the measures must be adapted to discrete data, e.g. annual changes.

These derivations are adapted to discrete, e.g. annual, data as shown in Table 2. The analytical derivations of TFP used Eq. (24) which defines a Divisia index as the starting point. The formulation in Table 2 approximates the Divisia index by the Törnqvist index but this is not the only possibility. The paper by ten Raa and Shestalova (2011) presents and explains the different possible approximations to the Divisia index for discrete data and outlines their properties. We illustrate in the example of the dual expenditure or cost function but the same ideas are applied to each of the other forms. Table 2 shows the TFP from the dual expenditure cost function for each component based on the estimated elasticity values for a panel data sample.

6 Estimation Issues 1: Between Country Differences: Do They Converge?

The idea of testing for convergence in the performance of countries over time has evolved from the literature on macroeconomic growth models.

Although economic historians have a long tradition of investigating national economic convergence, Baumol (1986) was one of the first papers by an economist to bring the topic to the forefront of economists' attention. Baumol's key empirical finding was relatively simple: he investigated the total productivity growth in GDP per labour-hour recorded in 16 major economies over the period 1870–1979 and regressed this against the productivity level measured in each country in 1870:

$$\text{Growth Rate } 1870-1979 = 5.25 - 0.75\,(\ln \text{GDP per work} - \text{hour } 1870) \quad (26)$$

Table 2 Index number calculations for TFP in the dual expenditure (cost) function case with panel data: $i = 1, \ldots, I$ and $t = 1, \ldots, T$

Component	Expression	Comment
SEC is: $\left(1 - 1/E^C\right)\boldsymbol{\eta}_{\mathbf{y}}'\dot{\mathbf{y}}$	$\frac{1}{2}\sum_{r=1}^{R}\left[\left(1 - 1/E_{it}^C\right)\eta_{y_r it} + \left(1 - 1/E_{it-1}^C\right)\eta_{y_r it-1}\right]\left(\ln y_{rit} - \ln y_{rit-1}\right)$	For one sector of the gross value-added measure of GDP; multiple outputs are assumed
AEC is: $\left(\mathbf{s} - \boldsymbol{\eta}_{\mathbf{w}}\right)'\dot{\mathbf{w}}$	$\frac{1}{2}\sum_{k=1}^{K}\left[\left(s_{kit} - \eta_{w_k it}\right) + \left(s_{kit-1} - \eta_{w_k it-1}\right)\right]\left(\ln w_{kit} - \ln w_{kit-1}\right)$	Weights are the divergence between actual and estimated optimal cost shares
EXC is: $-\boldsymbol{\eta}_{\mathbf{z}}'\dot{\mathbf{z}}$	$\frac{1}{2}\sum_{l=1}^{L}\left[\left(-\eta_{z_l it}\right) + \left(-\eta_{z_l it-1}\right)\right]\left(\ln z_{lit} - \ln z_{lit-1}\right)$	Exogenous variables could include quasi-fixed inputs in short run
TC is: $-\eta_t$	$-\frac{1}{2}\left[\left(\partial \ln c_{it}^j/\partial t\right) + \left(\partial \ln c_{it-1}^j/\partial t\right)\right]$	Technological progress shifts the cost function for sector j downwards

With an $R^2 = 0.88$ Baumol concluded that the lower the starting level of productivity in a given country the higher was its subsequent rate of growth. In other words, unproductive economies caught up with the productivity leaders over a long period of time. However, Baumol also demonstrated that the catch-up effect was more pronounced in a cluster of market-orientated economies than in a cluster of centrally planned economies, and the catch-up or convergence factor was absent in a cluster of less developed economies. Subsequently, this empirical regularity was addressed theoretically and empirically by many other economists notably Barro and Sala-i-Martin (2004), who regress a model which states that the average growth rate of per capita real output y_{it} in country $i, i = 1 \ldots N$ over a fixed period depends negatively on the starting value $y_{i,t-T}$ and also depends on other variables, $\mathbf{z}'_{it}$:

$$(1/T) \ln \left(y_{it}/y_{i,t-T}\right) = a - \left(\left(1 - \mathrm{e}^{\beta t}\right)/T\right) \left(\ln y_{i,t-T}\right) + \mathbf{z}'_{it}\boldsymbol{\mu} + u_{it} \quad (27)$$

They demonstrate that with a constant saving rate, the Solow-Swan theoretical one-sector growth model gives:

$$\beta = (1 - \alpha)(n + x + \delta) \quad (28)$$

α elasticity of output with respect to capital in the Cobb-Douglas production function
n rate of population growth
x rate of labour-augmenting technical progress, i.e. the steady-state growth rate of output per capita, which we met earlier as Solow's residual measure of productivity
δ rate of depreciation of the capital stock.

Therefore, the log of income per effective worker is a weighted average of the initial value and the steady-state value of income per effective worker, with the weight on the initial value declining exponentially at the rate β. Barro and Sala-i-Martin referred to this finding as beta-convergence. Subsequently, a different form of convergence was also identified, e.g. as stated in Young et al. (2008): "when the dispersion of real per capita income across a group of countries falls over time there is sigma convergence; when the partial correlation between the growth in income over time and its initial level is negative there is beta-convergence". Barro and Sala-i-Martin sum up the debate by stating:

Two concepts of convergence are:

(i) a poorer country tends to grow faster than a rich one, (beta-convergence) (i.e. the transition growth rate to the steady state is higher the lower the initial value of output per capita)

and

(ii) the dispersion of income per capita across countries diminishes over time (sigma-convergence). They suggest: beta-convergence can lead to sigma-convergence but new disturbances appear which offset this effect.

These ideas have been carried over to the literature on efficiency and productivity analysis of national economies, by incorporating into the analysis the measured data envelopment analysis or stochastic frontier analysis efficiency scores. Beta-convergence is measured by regressing the change in the log of countrywide mean efficiency against the previous level of the log mean efficiency and the lagged log change. Beta-convergence occurs if the coefficient on the lagged level is negative. Sigma-convergence is measured by regressing the change in the deviation in the log of countrywide mean efficiency from the log of the whole sample mean efficiency against the lagged value and the lagged change of this deviation. Sigma-convergence is said to occur if the coefficient on the lagged value of the deviation is negative. Panel least squares and GMM estimation are usually used. This type of analysis has typically been found in studies of national banking and financial systems, e.g. Casu and Girardone (2010).

However, a very different approach to the idea of convergence of TFP emerges from the development of endogenous growth theory in the 1990s. To understand how this relates to the neoclassical theory that we have discussed so far it is useful to go back to the simplest aggregate production function.

As we saw in Sect. 2, TFP in much of the literature is measured as a residual between an index of outputs with weights summing to one and an index of inputs with weights summing to one as well. Such measures are defined by the standard neoclassical production function relating the aggregate output Y_{it} of country $i, i = 1 \ldots N$ at time $t, t = 1 \ldots T$ to its inputs of capital, K_{it} labour L_{it} and time. The impact of technological progress is contained in the role of the time variable which smoothly improves the production function as time passes

$$Y_{it} = f(K_{it}, L_{it}, t) \tag{29}$$

Usually, an explicit assumption about the impact of technological progress would be made, and in the standard neoclassical growth model developed by Solow (1956) and Swan (1956) this took the form of labour enhancing technical change:

$$Y_{it} = f\left(K_{it}, L_{i0}\mathrm{e}^{\rho_i t}\right) \tag{30}$$

This form of the production function, when assumptions of positive but diminishing marginal products of the inputs and constant returns to scale are imposed in conjunction with the standard aggregate demand constraint, the definition of net investment as gross capital formation less depreciation of the capital stock and the Keynesian investment savings equilibrium condition, leads to an equilibrium in which income per capita grows at the constant rate ρ_i, and the capital income ratio and the consumption income ratios are constant. Eventually poorer countries would catch up with richer countries, and if there were international differences in ρ_i these would be unexplained since ρ_i is assumed to arise exogenously from a black box. In particular, only two explanations existed for the improvement in performance by different countries: either some had higher rates of input accumulation, especially capital, or some had faster trends in the productivity of labour. Initial research suggested that the second factor accounted for most US growth in the first half of the twentieth century, while considerable evidence (Krugman 1994; Young 1995) favoured the first factor in the growth of the tiger-economies of Southeast Asia in the second half of the twentieth century.

There emerged from this literature a set of "stylized facts" about productivity growth, as described in Jones and Romer (2010).

(1) Labour productivity has grown at a sustained rate.
(2) Capital per worker has also grown at a sustained rate.
(3) The real interest rate, or return on capital, has been stable.
(4) The ratio of capital to output has also been stable.
(5) Capital and labour have captured stable shares of national income.
(6) Amongst the fast-growing countries of the world, there is an appreciable variation in the rate of growth "of the order of 2–5 percent".

In terms of the neoclassical growth model, the first five facts are predicted and fact (6) is left unexplained, it is simply the Solow residual which implies that growth arises in a country exogenously, who knows from where?

The importance of this from our point of view in comparing international economic performance is that Eq. (29) is the standard starting point for a very large part of the efficiency and productivity analysis literature. Technical change is identified with the passage of time and is usually assumed to be an exogenous factor in the estimation model, which seems to indicate no policy direction which could improve a country's prospects. However, this is very much at odds with subsequent developments in the macroeconomic productivity growth literature, leading Jones and Romer (2010) to define a new set of stylised facts appropriate to modern developments. In brief, these are:

(1) Increases in the extent of the market. Increased flows of goods, ideas, finance and people—via globalisation, as well as urbanisation—have increased the extent of the market for all workers and consumers.
(2) Accelerating growth. For thousands of years, growth in both population and per capita GDP has accelerated, rising from virtually zero to the relatively rapid rates observed in the last century.
(3) Variation in modern growth rates. The variation in the rate of growth of per capita GDP increases with the distance from the technology frontier.
(4) Large income and total factor productivity (TFP) differences. Differences in measured inputs explain less than half of the enormous cross-country differences in per capita GDP.
(5) Increases in human capital per worker. Human capital per worker is rising dramatically throughout the world.
(6) Long-run stability of relative wages. The rising quantity of human capital, relative to unskilled labour, has not been matched by a sustained decline in its relative price.

It is fact (3) which started the trend towards endogenous growth models and was originally noted by Romer (1986) who plotted the annual average growth rate of GDP per capita over the period 1960–1985 for a large number of developing economies[12] against the income per capita in 1960 relative to the USA. The USA defined the technology frontier when the countries started growing and those with the lowest GDP per capita relative to the USA in 1960 subsequently showed a much larger variation in annual growth rates than countries that started from a position closer to the USA.

[12]Recall that the Baumol (1986) did not find convergence for countries outside a small sample of the most developed economies. Romer's finding of large variation in TFP rates for different countries used a much larger sample of chiefly developing countries.

This led to the suggestion that it must be the behaviour of producers, consumers and policy makers in different countries that had the largest impact on the variations in national growth rates. Globalisation, urbanisation and human capital provision are now key factors in determining different rates of productivity growth and economic performance. This is both an incentive and an obstacle to efficiency and productivity analysis. It provides an incentive because the ability of efficiency and productivity analysis to incorporate different approaches and variables for modelling productivity growth is its main strength but the obstacle is that formulating a theory of production on these lines that can be summarised in an aggregate production function is very difficult. In particular, it is necessary to ensure that the return to capital including human capital does not diminish as capital is accumulated.

One way of thinking about this is shown by Romer (1994) and Stiroh (2001). Compare Eq. (29) above with the production function represented in Eq. (31) below:

$$Y_{it} = A(R)f(K_{it}, L_{it}, R_{it}) \tag{31}$$

In this equation, the new variable R is the stock of knowledge and ideas which may be partially embodied in human capital. Each country's output depends on its own stock of knowledge and ideas, R_{it}, but the production function shifts up over time because of the global stock of knowledge and ideas, R. It is the stock of knowledge that permits the non-diminishing returns to investment that mean that growth is not exogenously limited but can be endogenously determined. The return to investment in knowledge broadly defined is given by the marginal product of knowledge:

$$\partial Y_{it}/\partial R_{it} = A(R)f_{R_{it}} + f(K_{it}, L_{it}, R_{it})A'(R)(\mathrm{d}R/\mathrm{d}R_{it}) \tag{32}$$

This can remain high even when $f_{R_{it}}$, the rate of return on the country's own knowledge stock for a constant state of global knowledge, tends to zero and it also incorporates a spillover term in the last expression. Spillovers and more generally the concept that ideas and knowledge are non-rival goods which are only partially or perhaps not at all excludable means that productivity measurement incorporating endogenous growth theory offers a very wide range of modelling design possibilities but these, for example, in the format of Eq. (31) may be difficult to incorporate into a standard efficiency and productivity analysis framework.

Nevertheless, there is a wide-ranging literature on spillover estimation particularly in the context of Leontief input-output analysis (I-O) which allows that commodities can be both intermediate inputs and final goods.

ten Raa and Wolff (2000) offer an interesting suggestion for spillover measurement in the context of this input-output approach. Commenting that usually spillover effects in each sector are measured by a weighted average of R&D in the sectors supplying intermediate inputs, ten Raa and Wolff instead suggest that spillover effects in an industry can be measured by TFP growth in its supplying sectors and they build up an analysis of interindustry spillovers that distinguishes four factors: autonomous growth, R&D in the sector in question, direct productivity spillovers using the direct input-output linkages between sectors to weight the supplying sectors' TFP growth rates and capital embodied spillovers using the investment coefficient of the supplying sector's capital to weight its productivity growth. ten Raa and Wolff (2000) then argue that productivity growth in a sector is counted in the sectors that trigger it. They find that for the I-O tables for the USA for 1958–1987, it is computers and office equipment and electronic components which are the sectors to which most productivity growth is imputed.[13]

It seems essential therefore to allow for the widest possible range of variables in explaining international differences in productivity and performance and Jones and Romer suggest that differences in institutions must be the fundamental source of the wide differences in growth rates; by institutions they mean a very wide range of different factors in each society and economy that should be incorporated into TFP models at the international level, in particular that weak and strong institutions affect the adoption and utilisation of ideas from leading nations on the frontier and that the potential for ideas to diffuse across nations amplifies the key role of institutions.

The role of institutions in TFP has been particularly strong in the work of North (1991) and Acemoglu et al. (2005) whose definition of good economic institutions means those that provide security of property rights and relatively equal access to economic resources to a broad cross section of society. The key argument here is the difference between the proximate causes of long-run TFP, i.e. factors like innovation and the spillover of ideas, which to North are not the causes of growth but are growth itself, and the fundamental causes of long-run TFP which are embedded in the evolution of society and the emergence of good economic institutions. This poses a problem for researchers: efficiency and productivity analysis by necessity focuses only on the proximate causes of TFP and, even then, the issues such as the form of

[13]The number of studies confirming the role of information technology in driving innovation and productivity growth throughout the world is expanding rapidly, see, for example, the long-term study in Chen and Fu (2018).

the production technology are difficult to model. Much more difficult is the problem of applying efficiency and productivity analysis to the understanding of the historical evolution of the fundamental causes of long-run TFPC.

The paper by Sickles et al. (2017) offers one solution to the problem of measuring what we must now call the proximate causes of TFP, i.e. the impact of innovation and new ideas in the production technology. In this treatment, Eq. (29) is the starting point, i.e. the essential neoclassical formulation of the production function. When this is extended to incorporate ideas of endogenous growth theory, Sickles et al. (2017) argue that the explanation for the spillover that endogenously determines technology change is the loosening of constraints on the utilisation of that technology, and that this is just another way of saying that TFP is primarily determined by the efficiency with which the existing technology (inclusive of innovations) is utilised.

Transformed into an empirical equation: write y_{it} as the log of GDP per capita in country $i.i = 1 \ldots N$ at time $t, t = 1 \ldots T$, write $\mathbf{X}_{it}$ as the vector of logged inputs and other technology factors including innovations some of which may be endogenous, and write $\eta_i(t)$ to represent the country-specific fixed effect, which may be time varying, so that with the error term v_{it}

$$y_{it} = \mathbf{X}'_{it}\beta + \eta_i(t) + v_{it} \tag{33}$$

$$v_{it} \sim Nid\left(0, \sigma_v^2\right) \tag{34}$$

This is the generic stochastic frontier analysis model of the production function. This is the basic model for estimating efficiency change using panel data frontier methods. If we assume that innovations are available to all countries and that idiosyncratic errors are due to relative inefficiencies, then the country-specific fixed effects can be used to capture the behavioural differences amongst countries that correspond to the key insight of the endogenous growth theory approach. Modern stochastic frontier analysis models offer a wide range of panel data methods for estimating the role of the countrywide time-varying fixed effects. The overall level of innovation change (innovation is assumed to be equally appropriable by all countries) can be measured directly by such factors as a distributed lag of R&D expenditures, or patent activity, or some such direct measure of innovation. In this way, Sickles et al. (2017) argue that the panel data methods incorporating endogeneity in the stochastic frontier analysis literature allow the researcher to address the issues raised by the endogenous growth models.

7 Estimation Issues 2: Technical Change

The technical change component of the decomposition of TFP is written as we saw as the log-derivative of the technology representation with respect to time, e.g. in the case of the production function in Table 1 using the Tornqvist form:

$$-\frac{1}{2}((\partial \ln y_{it}/\partial t) + (\partial \ln y_{it-1}/\partial t)) \tag{35}$$

The convention is to construct a very general role for the technical change component so that it shifts the whole production function (cost function) upwards (downwards) as illustrated in Fig. 1a. The nature of the technical change may be classified as Hicks-neutral if the ratio of marginal products of two inputs remains unchanged when the ratio of the inputs is unchanged, or the technical change may be labour-augmenting or capital augmenting (Harrod neutral or Solow neutral). In the case of a Cobb-Douglas production function, all three forms of technical change have the same parametric form but the measured rate of technical change in the labour-augmenting (capital augmenting) case is the Hicks-neutral rate scaled down by the output elasticity of labour (capital). Alternatively, the technical change may be non-neutral in which case it will depend on the levels of the inputs and possibly other variables as well.

However, in an important but to some extent empirically neglected paper, Atkinson and Stiglitz (1969) discussed the idea of localised technical change, as illustrated in Fig. 1b. Here, the smooth production function of Fig. 1a is a limiting case of the piecewise linear production function arising in the activity analysis approach to representing technology. This is the approach that also underlies the concept of the efficient frontier in data envelopment analysis. Technical change may then apply to a subset only of the portfolio of blueprint techniques available to producers. Atkinson and Stiglitz give the appealing example of a technical change in textile production that applies to a single technique rather than to every technique from a fully automated loom to the crudest handloom. The type of technical change which lifts the whole production function implies that technical progress spills over to every technique in the portfolio of technology. Localised technical change on the other hand limits the potential for spillovers from gains in knowledge from one form of production to another. The nature of the technical change is important here—the digital and information revolution may have much greater spillover potential for all techniques than, for example, the types of

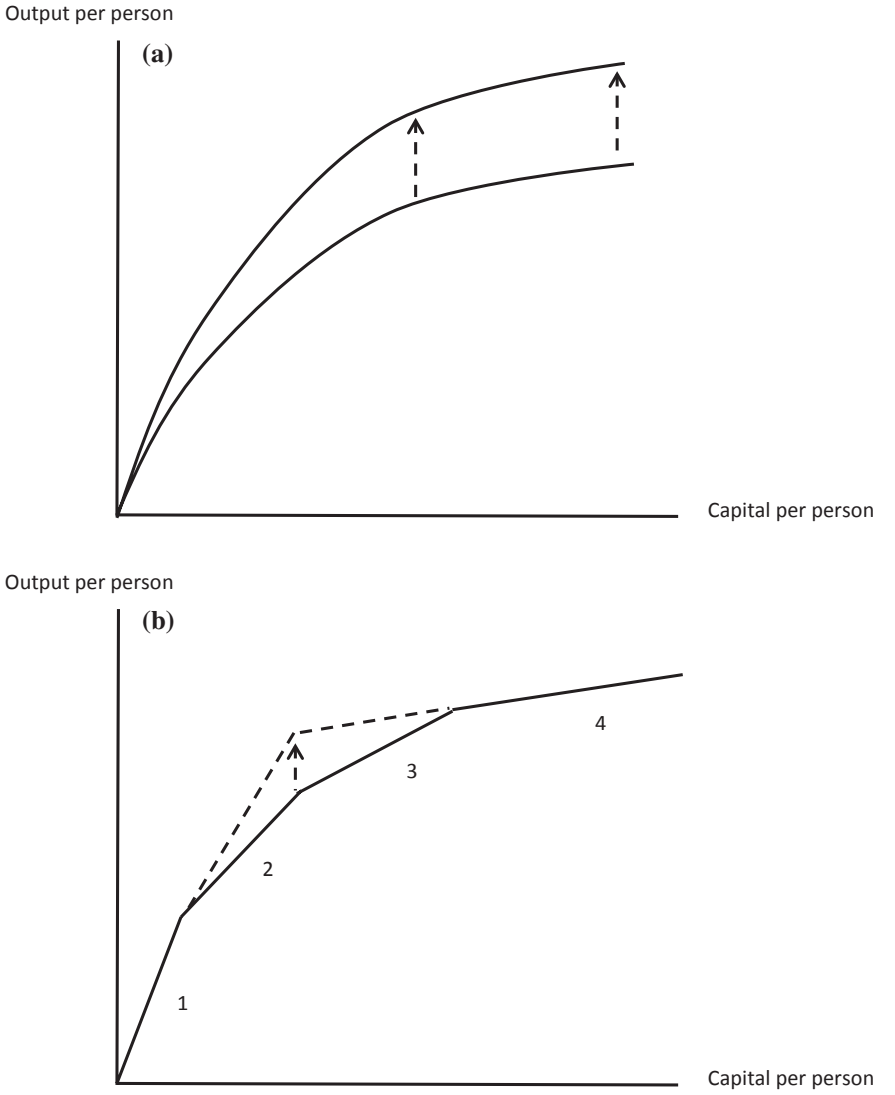

Fig. 1 **a** Conventionally representing the role of technical change as a shift in the whole production function. **b** Representing the role of technical change as a shift in a localised technique of production, Atkinson and Stiglitz (1969)

specific technological innovations which Gordon (2016) argues constituted the industrial revolution.

Two empirical issues arise from this concept of localised technical change, in addition to requiring a re-evaluation of the potential for spillovers. The first is that a technical improvement in a specific technique may have an

effect on the piecewise linear production function that means other techniques are dropped from the portfolio of technologies. In Fig. 1b, we can see that a localised technical change in the marginal product of technique 2 only, shifts the corresponding segment of the piecewise linear production frontier so that technique 3 will no longer be relevant to production as the scale expands; production moves directly from technique 2 to technique 4. This has the added effect that improvements in a localised technology that result in potential techniques dropping out of the portfolio may reduce the elasticity of substitution between inputs.

The second issue is empirical. Modelling of the possibility of technical change and its part in the decomposition of total factor productivity change may require non-parametric or semi-parametric estimation techniques if the existence of localised technical change makes the usual parametric functions (in which technical change shifts the whole function) inappropriate for the data-generation process.

This idea has had a consistent following in the empirical literature since it first appeared and amongst the most recent treatments is Acemoglu (2015) where the link is made with induced technical change and developments in *localised, biased and directed* technological change. For example, Acemoglu points out that frontier technologies developed in rich, capital-intensive countries may be inappropriate to a capital-scarce developing economy where such machinery may be limited. This approach has however had limited impact yet on the efficiency and productivity analysis modelling that we have been surveying here, and it poses problems for the way in which stochastic frontier analysis models can be specified.

8 Estimation Issues 3: Spillovers and Spatial Effects

In further research, one can consider similarities across countries since the performance of one national economy cannot be easily separated from that of its closest neighbours to assess how efficiency and productivity analysis using the newest developments in spatial econometrics can contribute to this question. In other words: Why might spatial spillovers be important in understanding the performance of national economies?

Spatial analysis in general has a long history in statistical modelling, and spatial econometrics has become recognised in recent years as an important new field of applications. In this survey of work on the performance of national economies, we do not have space to give a full summary of spatial

econometric applications but we can indicate briefly a relatively recent development which is the specification of a spatial econometric model with stochastic frontier analysis. We do this because there is a small emerging literature on using this approach to begin to understand the role of spatial spillovers on the performance of national economies with stochastic frontier analysis. Consider the aggregate production function relating output y to inputs $\mathbf{x}$, other exogenous variables $\mathbf{z}$ and time t together with the usual composed error term from stochastic frontier analysis as shown in Table 1:

$$\ln y = \ln f\left(\mathbf{x}', \mathbf{z}', t\right) - u + v \tag{36}$$

Conventionally, this is fitted as a Cobb-Douglas or translog functional form, but at present there is no allowance for spillovers onto the technology of one country from the technological advances in another neighbouring country. Spatial econometrics repairs this gap and we can make a simple start with the following Cobb-Douglas specification adapted from Glass et al. (2016) for a cross section of countries labelled $i$or j over time periods labelled t.

$$\ln y_{it} = \alpha + \sum_k \beta_k \ln x_{kit} + \sum_l \gamma_l \ln z_{lit} + \rho t + \delta \sum_{j \neq i} w_{ij} \ln y_{jt} - u_{it} + v_{it} \tag{37}$$

This production function is very familiar in the first three expressions representing inputs and other exogenous variables with constant output elasticities together with Hicks-neutral technological progress. The fourth expression however adds a weighted summation of the output levels in other countries to the explanation of production in country i. This spatial autoregressive model SAR represents the effect of accumulated spatial lags which are one way of modelling spillovers from one technology to another. With appropriate numerical values for the spatial weights matrix $\left[w_{ij}\right] = \mathbf{W}$ we can devise a set of explanatory variables whose spillover effects are captured by the estimated parameter δ. These imposed numerical values permit the researcher to investigate a variety of nearby neighbour effects based on geographical dispersion to capture possible spillovers. A multiplicity of extensions to this idea can be devised, see, for example, Greene (2017), which include the application of the spatial weights matrix to the explanatory variables and to the error components.

There are issues of interpretation of the results that require careful analysis, as Glass et al. (2016) point out, "a unit in a spatial model is therefore simultaneously exporting and importing spillovers to and from its neighbours. The indirect marginal effects from a spatial model measure the magnitude of the spillovers which are imported and exported in the sample". More interesting from our point of view of the estimation issues in the

modelling of national performance by stochastic frontier analysis is how the composed error term can be addressed. Use can be made of a concentrated likelihood function approach first suggested by Fan et al. (1996). This was used in the spatial autoregressive model by Glass et al. (2016) and its use in non-parametric estimation is recommended by Kuosmanen et al. (2015). We can show here the Glass et al. (2016) procedure adapted to the problem in hand.

The first-order conditions for the log-likelihood function for the stochastic frontier analysis model are still valid even if the frontier is unknown and estimated separately, provided it does not depend on $\sigma^2 = \sigma_v^2 + \sigma_u^2$ and $\lambda = \sigma_u/\sigma_v$.

Following Glass et al. (2016), a two-step procedure is available:

Step 1 Solve the spatial regression model estimators retaining the SAR residuals

Step 2 Now use these to obtain the concentrated log-likelihood in terms of $\lambda = \sigma_u/\sigma_v$ only. Maximise this by grid search for $\widehat{\lambda}$ and iterate jointly with $\widehat{\sigma}^2$ to convergence.

With the estimators obtained, a transformation of the usual measures of conditional efficiency can be derived and these depend on turn on the spatial lag effects; these results can then be written as direct, indirect and total efficiency measures. Glass et al. (2016) used this approach to estimate a stochastic frontier analysis aggregate production function using aggregate data for 41 European countries for the period 1990–2011 with a dense spatial weights matrix based on distances. The output variable was real aggregate value added, and the inputs were capital and labour with additional variables of export openness relative to GDP and government expenditure relative to GDP. A key finding was that on average, countries are more adept at importing efficiency than they are at exporting efficiency. This finding is consistent with the diffusion of knowledge embodied in imports of hi-tech goods and services from a relatively small number of technological leaders in the sample (e.g. Germany).

9 Summary and Conclusions

This chapter serves as an introduction to the issues of comparing the performance of national economies. We could not hope to survey in detail the massive number of empirical papers that have accumulated on this topic and of course the methodologies of data envelopment analysis and stochastic

frontier analysis are well covered in other chapters; therefore, we could have little to add on strict methodology. Instead, we have opted to present a broad overview of a wide range of different topics of relevance to the general idea of comparing the national performance of countries.

We began with basic historical ideas that are still important for researchers coming new to the topic. Productivity comparisons are critical and are made every day in the media and in political and economic commentary. We showed how a myriad of different ideas have evolved from the original growth model of Solow and the identification of TFP with a residual. We questioned whether the key variable of real value-added GDP tells us anything about economic welfare and presented a few ideas on this from the current wealth of contributions that are available, including the suggestion of measuring the "happiness frontier". Then we investigated the roles of data envelopment analysis and stochastic frontier analysis in making efficiency and productivity comparisons amongst countries. Our key argument here is that the data envelopment analysis and stochastic frontier analysis approaches permit the relaxation of the major assumptions associated with TFP measures reported in the media and which are usually the basis of policy making. We explored ways in which the data envelopment analysis and the stochastic frontier analysis permit the researcher to relax the assumptions of allocative efficiency, constant returns to scale, absence of exogenous variable effects (the *ceteris paribus* assumption) and the absence of inefficient performance that characterise the conventional growth accounting or neoclassical approach to the comparison of national economic performance.

From there we investigated a number of estimation issues both settled and unsettled in the efficiency and productivity analysis approach to national TFP measurement.

We considered the ideas about convergence of national performance and how this might be measured, and we saw the contrast between the convergence in national economic performance and TFP rates predicted by the neoclassical model and the lack of convergence due to the endogeneity innovations associated with the endogenous growth model. A second estimation issue concerned the modelling of technological change and whether this applied to the whole representation of the frontier as is conventional in stochastic frontier analysis or whether we could consider localised technical change as initially suggested by Atkinson and Stiglitz. Intuitively, it seemed as if data envelopment analysis or other non-parametric approaches could be more fruitful than stochastic frontier analysis in this context but researcher ingenuity will no doubt overcome this. The final estimation issue that we examined was the interface between spatial econometrics and stochastic

frontier analysis and we gave an example of comparison of national economic performance in which the composed error term of stochastic frontier analysis was incorporated in a spatial autoregressive model using a concentrated likelihood estimation approach.

In many ways, this chapter differs from the other technical chapters in this book. However, this is deliberate. Our intention has been to provide a broad overview of the whole context in which we can compare, as economists particularly interested in efficiency and productivity analysis, the performance of national economies. We have deliberately not attempted the impossible task of summarising the empirical literature on international differences in TFP, even those using only efficiency and productivity analysis since there are literally tens of thousands of such papers. Instead, we have consciously taken a wide and eclectic view about what constitutes international economic performance, in the belief that the powerful tools of efficiency and productivity analysis will successfully address these massive issues gaining an accurate picture of how different countries compare with each other using the widest range of concepts of what constitutes a country's economic performance.

References

Acemoglu, D., S. Johnson, and J.A. Robinson. 2005. Institutions as a fundamental cause of long-run growth. In *Handbook of economic growth*, vol. 1, ed. Philippe Aghion and Steven N. Durlauf, 385–472. Amsterdam: Elsevier B.V.

Acemoglu, Daron. 2015. Localised and biased technologies: Atkinson and Stiglitz's new view, induced innovations, and directed technological change. *The Economic Journal* 125 (583): 443–463.

Aigner, D., C.K. Lovell, and P. Schmidt. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6 (1): 21–37.

Atkinson, Anthony B., and Joseph E. Stiglitz. 1969, September. A new view of technological change. *The Economic Journal* 79 (315): 573–578.

Baker, Dean, J. Bradford DeLong, and Paul Krugman. 2005. Asset returns and economic growth. *Brookings Papers on Economic Activity* 2005 (1): 289–330.

Barro R., and X. Sala-i-Martin. 2004. *Economic growth*. Cambridge: MIT Press.

Bauer P. 1990. Recent developments in the econometric estimation of frontiers. *Journal of Econometrics* 46 (1–2): 39–56.

Baumol, William J., Robert E. Litan, and Carl J. Schramm. 2007. *Good capitalism, bad capitalism, and the economics of growth and prosperity*. New Haven: Yale University Press.

Baumol, William J. 1986. Productivity growth, convergence, and welfare: What the long-run data show. *The American Economic Review* 1986: 1072–1085.

Bettio, Francesca, and Janneke Plantenga. 2004. Comparing care regimes in Europe. *Feminist Economics* 10 (1): 85–113. https://doi.org/10.1080/1354570042000198245.

Bogetoft, P., and L. Otto. 2011. *Benchmarking with DEA, SFA, and R*. New York: Springer.

Casu, B., and C. Girardone. 2010. Integration and efficiency convergence in EU banking markets. *Omega* 38 (5): 260–267.

Caves, D.W., L.R. Christensen, and W.E. Diewert. 1982. The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica: Journal of the Econometric Society* 50 (6): 1393–1414.

Charnes, A., W.W. Cooper, and E. Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2 (6): 429–444.

Chen, Hao-Tsung, and Tsu-Tan Fu. 2018. Source of growth analysis at the industry-level for selected Asian economies using DEA Malmquist and Asia KLEMS data bases. Asia-Pacific Productivity Conference, APPC2018 at Seoul University, Korea, July 2018.

Coelli T, A. Estache, S. Perelman, and L. Trujillo. 2003. *A primer on efficiency measurement for utilities and transport regulators*. Washington, DC: World Bank.

Coyle, Diane. 2014. *GDP: A brief but affectionate history*. Princeton, NJ: Princeton University Press.

Coyle, Diane. 2017. Do-it-yourself digital: the production boundary and the productivity puzzle. ESCoE Discussion Paper 2017-01, Economic Statistics Centre of Excellence, June 2017.

Diewert, W.E. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4 (2): 115–145.

Easterlin, R. 1974. Does economic growth improve the human lot? Some empirical evidence. In *Nations and households in economic growth: Essays in honor of Moses Abramovitz*, ed. P.A. David and M.W. Reder, 89–125. New York: Academic Press.

Fan, Yanqin, Qi Li, and Alfons Weersink. 1996. Semiparametric estimation of stochastic production frontier models. *Journal of Business & Economic Statistics* 14 (4): 460–468.

Färe, R., S. Grosskopf, M. Norris, and Z. Zhang. 1994. Productivity growth, technical progress, and efficiency change in industrialized countries. *American Economic Review* 84: 66–83.

Färe, R., and D. Primont. 2012. *Multi-output production and duality: Theory and applications*. New York: Springer Science & Business Media.

Farrell, Michael. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society* 120: 253–281.

Folbre, Nancy, and Julie A. Nelson. 2000. For love or money—Or both. *The Journal of Economic Perspectives* 14 (4, Autumn): 123–140.

Giraleas, Dimitris, Ali Emrouznejad, and Emmanuel Thanassoulis. 2012. Productivity change using growth accounting and frontier-based approaches—Evidence from a Monte Carlo analysis. *European Journal of Operational Research* 222 (3): 673–683.

Glass, Anthony J., Karligash Kenjegalieva, and Robin C. Sickles. 2016. A spatial autoregressive stochastic frontier model for panel data with asymmetric efficiency spillovers. *Journal of Econometrics* 190 (2): 289–300.

Goodridge, Peter, Jonathan Haskel, and Gavin Wallis. 2016, December. Accounting for the UK productivity puzzle: A decomposition and predictions. *Economica*. https://doi.org/10.1111/ecca.12219.

Gordon, R.J. 2003. Exploding productivity growth: Context, causes, and implications. *Brookings Papers on Economic Activity* 2003 (2): 207–298.

Gordon, Robert J. 2016. *The rise and fall of American growth: The U.S. standard of living since the civil war*. Princeton: Princeton University Press.

Greene, William H. (2017) *Econometric analysis*, 8th ed. New York: Pearson.

Haskel, Jonathan, and Stian Westlake. 2017. *Capitalism without capital: The rise of the intangible economy*. Princeton: Princeton University Press.

Helliwell, John, Richard Layard, and Jeffrey Sachs (eds.). 2012. *World happiness report*. New York, NY: The Earth Institute, Columbia University.

Jones, C.I., and P.M. Romer. 2010. The new Kaldor facts: Ideas, institutions, population, and human capital. *American Economic Journal: Macroeconomics* 2 (1): 224–245.

Jones, Charles I., and Peter J. Klenow. 2016. Beyond GDP? Welfare across countries and time. *American Economic Review* 106 (9): 2426–2457.

Jorgenson, Dale, and Zvi Griliches. 1967. The explanation of productivity change. *The Review of Economic Studies* 34 (3): 249–283.

Kuosmanen, Timo, Andrew Johnson, and Antti Saastamoinen. 2015. Stochastic nonparametric approach to efficiency analysis: A unified framework. In *Data envelopment analysis*, 191–244. Boston, MA: Springer.

Krugman, P. 1994. The myth of Asia's miracle. *Foreign Affairs* 73: 62–78.

Lovell, C.A. Knox. 2003. The decomposition of Malmquist productivity indexes. *Journal of Productivity Analysis* 20: 437–458.

Lukasz Rachel, and Thomas D. Smith. 2015. Secular drivers of the global real interest rate. Staff Working Paper No. 571, Bank of England.

Milner, C., and T. Weyman-Jones. 2003. Relative national efficiency and country size: Evidence for developing countries. *Review of Development Economics* 7 (1): 1–14.

North, D.C. 1991. Institutions. *Journal of Economic Perspectives* 5 (1): 97–112.

Orea, L. 2002. Parametric decomposition of a generalized Malmquist productivity index. *Journal of Productivity Analysis* 18: 5–22.

Panzar, J.C., and R.D. Willig. 1977. Economies of scale in multi-output production. *The Quarterly Journal of Economics* 91: 481–493.

Ray, S.C., and E. Desli. 1997. Productivity growth, technical progress, and efficiency change in industrialized countries: Comment. *The American Economic Review* 87 (5): 1033–1039.

Romer, P.M. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94 (5): 1002–1037.

Romer, P.M. 1994. The origins of endogenous growth. *Journal of Economic Perspectives* 8 (1): 3–22.

Sickles, Robin C., Jiaqi Hao, and Chenjun Shang. 2017. *Panel data and productivity measurement Chapter 17*. In *Oxford handbook of panel data*, ed. Badi Baltagi. New York: Oxford University Press.

Simar, L., and P.W. Wilson. 2007. Estimation and inference in two-stage semi-parametric models of production processes. *Journal of Econometrics* 136: 31–64.

Skidelsky, Robert. 2003. *John Maynard Keynes 1883–1946: Economist, philosopher, statesman*. London: Macmillan.

Solow, R.M. 1956. A contribution to the theory of economic growth. *The Quarterly Journal of Economics* 70 (1): 65–94.

Solow, R.M. 1957. Technical change and the aggregate production function. *The Review of Economics and Statistics* 39 (3): 312–320.

Smith, Adam. 1776. *An inquiry into the nature and causes of the wealth of nations*, ed. R.H. Campbell, A.S. Skinner, and W.B. Todd, 2 vols., Indianapolis: Liberty Fund 1981.

Stiroh, K.J. 2001, March. What drives productivity growth? *Economic Policy Review* 7: 37, Federal Reserve Bank of New York.

Summers, Robert, and Alan Heston. 1991, May. The Penn World Table (Mark 5): An expanded set of international comparisons, 1950–1988. *Quarterly Journal of Economics* 106 (2): 327–368.

Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32 (2): 334–361.

ten Raa, T., and P. Mohnen. 2002. Neoclassical growth accounting and frontier analysis: A synthesis. *Journal of Productivity Analysis* 18 (2): 111–128.

ten Raa, T., and V. Shestalova. 2011. The Solow residual, Domar aggregation, and inefficiency: A synthesis of TFP measures. *Journal of Productivity Analysis* 36 (1): 71–77.

ten Raa, T., and E.N. Wolff. 2000. Engines of growth in the US economy. *Structural Change and Economic Dynamics* 11 (4): 473–489.

Thaler, R.H. 2018. From cashews to nudges: The evolution of behavioral economics. *American Economic Review* 108 (6): 1265–1287.

Varian, Hal. 2016. *A microeconomist looks at productivity: A view from the valley*. The Brookings Institution. https://www.brookings.edu/wp-content/uploads/2016/08/varian.pdf. Accessed February 2018.

Winsten, C.B. 1957. Discussion on Mr. Farrell's paper. *Journal of the Royal Statistical Society* 120 (3): 282–284.

Young, Andrew T., Matthew J. Higgins, and Daniel Levy. 2008. Sigma convergence versus beta convergence: Evidence from US county—Level data. *Journal of Money, Credit and Banking* 40 (5): 1083–1093.

Young, A. 1995. The tyranny of numbers: Confronting the statistical realities of the East Asian growth experience. *The Quarterly Journal of Economics* 110 (3): 641–680.

Productivity Indexes and National Statistics: Theory, Methods and Challenges

W. Erwin Diewert and Kevin J. Fox

> … it is not reasonable for us to expect the government to produce statistics in areas where concepts are mushy and where there is little professional agreement on what is to be measured and how. (Griliches 1994 Presidential Address to the American Economic Association, p. 14)

1 Introduction

Productivity is a major driver of long-term economic growth and welfare improvements. Productivity indexes are used in a wide variety of policy contexts, such as for government budget forecasting, designing innovation policy and assessing the relative effectiveness of government policies.

Productivity growth slowdowns cause much policy debate and concern. The slowdown from the early 1970s to the mid-1990s in many industrialized countries was much debated, especially given that this was the period during which personal computers diffused rapidly into workplaces. This

W. E. Diewert
Vancouver School of Economics,
University of British Columbia, Vancouver, BC, Canada
e-mail: erwin.diewert@ubc.ca

W. E. Diewert · K. J. Fox (✉)
School of Economics, UNSW, Sydney, NSW, Australia
e-mail: K.Fox@unsw.edu.au

© The Author(s) 2019
T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_18

resulted in much attention to the measurement of productivity—if the slowdown was simply a case of measurement lagging behind developments in an increasingly complex economy, then the solution is to modernize the collection and construction of economic statistics (see Diewert and Fox [1999] and references therein for more on this slow growth episode and potential explanations).

From around the mid-1990s, a measured increase in productivity allayed concerns and was viewed in most countries as the benefits of computerization finally being realized. However, a subsequent decline in productivity growth since 2004 across all industrialized countries has again heightened concerns. It raises the possibility that the earlier productivity recovery was an unusual episode and that lower growth is the new norm. In the latter case, we should expect lower increases in living standards in the future. This is the view of, for example, Gordon (2016) and Cowen (2011).[1] Alternatively, perhaps measurement problems associated with the digital economy and rapid quality change in products have come to the fore. In his Independent Review of UK Economic Statistics, interim report, Bean (2016, p. 7) noted that "Statistics have failed to keep pace with the impact of digital technology". This concern has yielded a growing literature on measurement problems for National Statistical Offices (NSOs), their potential to explain away the productivity slowdown and alternative approaches to measuring economic activity in a modern economy.[2]

In this chapter, to provide a means to better understand such debates, we begin by examining the theoretical basics of productivity growth measurement as employed by NSOs. In particular, we provide the theoretical justifications for the index number formulae that are commonly used. We then turn to a discussion of data used in index number construction in practice and highlight the measurement challenges.

The productivity of a production unit is defined as the output produced by the unit divided by the input used over the same time period.[3] If the input measure is comprehensive, then the productivity concept is

[1]However, others provide a more optimistic view; see, e.g., Sichel (2016), Mokyr et al. (2015) and Brynjolfsson and McAfee (2011, 2014).

[2]See, for example, Brynjolfsson et al. (2019), IMF (2018), Diewert et al. (2018), Feldstein (2017), Groshen et al. (2017), Hulten and Nakamura (2017), Syverson (2017), Ahmad and Schreyer (2016), Byrne et al.(2016), Brynjolfsson and Saunders (2009), Brynjolfsson and Oh (2012) and Greenstein and McDevitt (2011).

[3]A production unit could be an establishment, a firm, an industry or an entire economy.

called *Total Factor Productivity* (*TFP*) or *Multifactor Productivity* (*MFP*).[4] If the input measure is labour hours, then the productivity concept is called *Labour Productivity*.

The Bureau of Labor Statistics in the USA was the first NSO to introduce an official program to measure *MFP* in 1983 (see Dean and Harper 2001). Other countries with *MFP* programs now include Canada, Australia, the UK and New Zealand. The OECD also publishes *MFP* and Labour Productivity statistics for member countries (see OECD 2018).

We will focus on *MFP* and how to measure it rather than Labour Productivity. The Labour Productivity concept has its uses but the problem with this concept is that it could be very high in one country compared to another country with the difference being entirely due to a larger amount of non-labour input in the first country. On the other hand, if *MFP* is much higher in country A compared to country B, then country A will be genuinely more efficient than country B and it will be useful to study the organization of production in country A in order to see if the techniques used there could be exported to less efficient countries.

A problem with the *MFP* concept is that it depends on the units of measurement for outputs and inputs. Hence, *MFP* can only be compared across production units if the production units are basically in the same line of business so that they are producing the same (or closely similar) outputs and using the same inputs. However, in the time series context, Multifactor Productivity *growth rates* can be compared over dissimilar production units, and hence, we will focus most of our attention on measuring Multifactor Productivity Growth (*MFPG*).

We begin by providing an introduction to the issues involved in measuring *MFPG* by considering the special case where the production unit produces only a single output and uses only a single input. It turns out in this case that there are four equivalent ways for measuring *MFPG*. Section 3 generalizes this framework to the multiple input and output case, as faced by NSOs. This requires the choice of index number formula. Section 4 examines this problem from the test (or "axiomatic") approach. Essentially, this involves comparing the mathematical properties of the formula against a battery of tests which are deemed as desirable. Section 5 examines this

[4]The terms "Multifactor Productivity" and "Total Factor Productivity" are synonymous. National Statistical Offices tend to favour "Multifactor Productivity", presumably to avoid giving the impression that a claim is being made that all factors of production have been taken into account. Academics are typically immune to such caution and tend to use the term "Total Factor Productivity". As our focus is on NSO practice, we will use the term "Multifactor Productivity".

formula choice problem from the perspective of economic theory, recognizing that the resulting indexes are measuring economic concepts. Thus, these sections provide the justification for the index number choices made by NSOs in constructing productivity growth estimates.

Section 6 discusses data needs for constructing the productivity indexes and reviews the concepts, sources and methods that are used for the output, labour and capital components. Section 7 highlights several difficult measurement problems faced by NSOs and suggests some ways forward. Section 8 concludes.

2 Productivity Measurement in the Case of One Input and One Output

We consider in this section the problem of measuring the Multifactor Productivity (*MFP*) (and the growth of Multifactor Productivity, *MFPG*) of a one output, one input firm.[5] To do this, we require data on the amounts of output produced, y^0 and y^1, during two time periods, 0 and 1, and on the amounts of input utilized, x^0 and x^1, during those same two time periods. It is also convenient to define the firm's revenues R^t and total costs C^t for period t where $t=0, 1$. The average selling price of a unit of output in period t is assumed to be p^t and the average cost of a unit of input in period t is w^t for $t=0, 1$. Thus, we have:

$$R^t = p^t y^t \quad \text{for } t = 0, 1 \tag{1}$$

and

$$C^t = w^t x^t \quad \text{for } t = 0, 1. \tag{2}$$

Our first definition of the *MFPG* of the firm going from period 0 to period 1 (or more briefly, of the productivity of the firm) is:

$$MFPG(1) = \left(y^1/y^0\right) / \left(x^1/x^0\right). \tag{3}$$

Note that y^1/y^0 is (one plus) the firm's output growth rate going from period 0 to period 1 while x^1/x^0 is the corresponding input growth rate going from

[5]The material in this section is largely taken from Diewert (1992) and Diewert and Nakamura (2003).

period 0 to period 1.[6] If $MFPG(1) > 1$, then the output growth rate was greater than the input growth rate and we say that the firm has experienced a *productivity improvement* going from period 0 to period 1. If $MFPG(1) < 1$, then we say that the firm has experienced a *productivity decline*.

The output growth rate, y^1/y^0, can also be interpreted as a *quantity index of outputs*. Indeed, in the following section where we consider the case of multiple outputs, we will replace y^1/y^0 by a quantity index for outputs. However, if there is only one output, it can be verified that the output quantity indexes defined there all reduce to the output growth rate, y^1/y^0. Similarly, the input growth rate, x^1/x^0, can be interpreted as a quantity index of inputs. Hence, our first definition of productivity growth, $MFPG(1)$ defined by (3), can be interpreted as an output quantity index divided by an input quantity index.

An alternative method for measuring productivity in a one output, one input firm is the *change in technical coefficients* method. Define the input-output coefficient of the firm in period t as:

$$a^t \equiv y^t/x^t, \quad t = 0, 1. \tag{4}$$

Thus, a^t is the total amount of output y^t produced by the firm in period t divided by the total amount of input utilized by the firm in period t, x^t. It can be interpreted as a coefficient which summarizes the engineering and economic characteristics of the firm's technology in period t: a^t describes the rate at which inputs are transformed into outputs during period t.

Our second definition of total factor productivity can be expressed in terms of the output-input coefficients, a^0 and a^1, as follows:

$$MFPG(2) = a^1/a^0. \tag{5}$$

Thus, if a^1 is greater than a^0, so that the firm is producing more output per unit input in period 1 compared to period 0, then $MFPG(2)$ and the firm has experienced an increase in productivity going from period 0 to period 1.

It should be noted that the two productivity growth concepts that we have defined thus far, $MFPG(1)$ and $MFPG(2)$, are both relative concepts. This is a general feature of economic definitions of productivity: the performance of the firm in a current period 1 is always compared to its performance in a base period 0. In contrast, an engineering concept of

[6] In what follows, we will somewhat incorrectly refer to y^1/y^0 as the output growth rate and x^1/x^0 as the input growth rate, where these are both actually one plus growth rates.

productivity or efficiency is usually an absolute one, concerned with obtaining the maximum amount of output in period one, y^1, given an available amount of input in period one, x^1, consistent with the laws of physics.[7]

Using (3), (4) and (5), it is easy to show that $MFPG(2)$ coincides with an earlier $MFPG(1)$ concept in this simple one output, one input model of production; i.e., we have:

$$\begin{aligned} MFPG(2) &= a^1/a^0 = \left(y^1/x^1\right)/\left(y^0/x^0\right) = \left(y^1/y^0\right)/\left(x^1/x^0\right) \\ &= MFPG(1). \end{aligned} \tag{6}$$

We turn now to a third possible method for defining productivity:

$$MFPG(3) \equiv \left[\left(R^1/R^0\right)/\left(p^1/p^0\right)\right]/\left[\left(C^1/C^0\right)/\left(w^1/w^0\right)\right]. \tag{7}$$

Thus, $MFPG(3)$ is equal to the firm's revenue ratio R^1/R^0 deflated by the output price index p^1/p^0 divided by the cost ratio between the two periods C^1/C^0 deflated by the input price index w^1/w^0.

Using (1), we have

$$\left(R^1/R^0\right)/\left(p^1/p^0\right) = \left(p^1y^1/p^0y^0\right)/\left(p^1/p^0\right) = y^1/y^0 \tag{8}$$

and using (2), we have

$$\left(C^1/C^0\right)/\left(w^1/w^0\right) = \left(w^1x^1/w^0x^0\right)/\left(w^1/w^0\right) = x^1/x^0. \tag{9}$$

[7]The engineers Norman and Bahiri (1972, p. 27) define productivity as the quotient obtained by dividing output by one of the factors of production. Since our simple model has only one factor of production, this engineering definition of productivity reduces to $a^1 = y^1/x^1$. However, even engineers recognize that this definition of productivity is unsatisfactory, since it is not invariant to changes in the units of measurement. Thus, Norman and Bahiri (1972, p. 28) later define productivity as a relative concept as the following quotation indicates:

> Consequently, we define and measure relative productivity levels in comparison with a level achieved in the past or in comparison with another establishment in the same industry, or in comparison with the national average achieved by another nation.

Thus, a^1 is compared to a^0 where $a^0 = y^0/x^0$ is a reference input-output coefficient. Note that a^1/a^0 is invariant to changes in the units of measurement. It should be mentioned that sometimes economists (such as Jorgenson and Griliches 1967, p. 252) define productivity as total output divided by total input, $y^1/x^1 = a^1$ and then define productivity change as the rate of change of a^1. However, it is only their productivity change concept that is regarded as being meaningful.

Thus, in this simple one input, one output model, (8) says that the deflated revenue ratio is equal to the output growth rate and (9) says that the deflated cost ratio is equal to the input growth rate. Hence, (7) equals (3) and we have, using (6):

$$MFPG(1) = MFPG(2) = MFPG(3). \tag{10}$$

There is a fourth way for measuring productivity change that is a generalization of a method originally suggested by Jorgenson and Griliches (1967). In order to explain this fourth method, we need to introduce the concept of the firm's period t margin, m^t; i.e., define

$$1 + m^t \equiv R^t/C^t; \quad t = 0, 1. \tag{11}$$

Thus, $1 + m^t$ is the ratio of the firm's period t revenues R^t to its period t costs C^t. If m^t is zero, then the firm's revenues equal its costs in period t and the economic profit of the firm is zero. If m^t is positive, then the bigger m^t is, the bigger are the firm's profits.

We can now define our fourth way for measuring productivity change in a one output, one input firm:

$$MFPG(4) \equiv \left[\left(1 + m^1\right)/\left(1 + m^0\right)\right]\left(w^1/w^0\right)/\left(p^1/p^0\right). \tag{12}$$

Thus, $MFPG(4)$ is equal to the margin growth rate $(1 + m^1)/(1 + m^0)$ times the rate of increase in input prices w^1/w^0 divided by the rate of increase in output prices p^1/p^0.

If we use Eqs. (11) to eliminate $(1 + m^1)/(1 + m^0)$ in (12), we find that

$$MFPG(4) = MFPG(3) \tag{13}$$

and thus, by (10), $MFPG(1) = MFPG(2) = MFPG(3) = MFPG(4)$. Thus, in a one output, one input firm, we have four conceptually distinct methods for measuring productivity change that turn out to be equivalent. Unfortunately, this equivalence does not generally extend to the multiple output, multiple input case.

Definition (12) of productivity can be used to show the importance of achieving a productivity gain: a productivity improvement is the source for increases in margins or increases in input prices or decreases in output prices. Equation (12) also indicates the relationship between total factor productivity and increased profitability. Rearranging (12), we have:

$$\left(1 + m^1\right)/\left(1 + m^0\right) = [MFPG(4)]\left(p^1/p^0\right)/\left(w^1/w^0\right). \tag{14}$$

Thus, the rate of growth in margins is equal to *MFPG* times the output price growth rate divided by the input price growth rate.

If there are constant returns to scale in production or margins m^t are zero for whatever reason in periods 0 and 1, then *MFPG*(4) reduces to $(w^1/w^0)/(p^1/p^0)$, which is the input price index divided by the output price index, a formula due to Jorgenson and Griliches (1967, p. 252).

We conclude this section with a rather lengthy discussion of the problem of distinguishing *MFPG* from the concept of technical change or technical progress, *TP*. In order to distinguish *MFPG* from *TP*, it is necessary to introduce the concept of the firm's period t production function f^t; i.e., in period t, $y=f^t(x)$ denotes the maximum amount of output y that can be produced by x units of the input. We assume that in periods 0 and 1, the observed amounts of output, y^0 and y^1, are produced by the observed amounts of input, x^0 and x^1, according to the following production function relationships:

$$y^0 = f^0\left(x^0\right); \tag{15}$$

$$y^1 = f^1\left(x^1\right). \tag{16}$$

Note that we are now explicitly assuming that production is technically efficient during the two periods under consideration.[8]

We define technical progress *TP* as a measure of the shift in the production function going from period 0 to period 1. There are an infinite number of possible shift measures but it turns out that four measures of technical progress (involving the observed data y^0, y^1, x^0 and x^1 in some way) are the most useful. First, define:

$$y^{0*} = f^1\left(x^0\right) \text{ and } y^{1*} = f^0\left(x^1\right). \tag{17}$$

[8]In benchmarking studies or in studies where we compare the relative efficiency of different production units producing the same outputs and using the same inputs, we do not assume that each production unit is globally efficient; i.e., the best practice production unit is regarded as being technically efficient but the other production units may not be technically efficient relative to the global best practice technology. In the time series context, it may be acceptable to assume that each production unit is technically efficient in each period *relative to its own knowledge of the technology available to it*. In other words, individual production units are efficient relative to their own knowledge base but of course they can be inefficient relative to the world wide best-practice technology.

Thus, y^{0*} is the output that could be produced by the period 0 input x^0 if the period 1 production function f^1 were available and y^{1*} is the output which could be produced by the period 1 input x^1 but using the period 0 technology which is summarized by the period 0 production function f^0. Note that in order to define these hypothetical outputs y^{0*} and y^{1*}, a knowledge of the period 0 and 1 production functions f^0 and f^1 is required. This knowledge is not easy to acquire but it could be obtained by engineering studies or by nonparametric or econometric methods for obtaining a suitable reference technology.

With y^{0*} and y^{1*} defined, we can define the following two *output-based indexes of technical progress* $TP(1)$ and $TP(2)$[9]:

$$TP(1) = y^{0*}/y^0 = f^1\left(x^0\right)/f^0\left(x^0\right); \tag{18}$$

$$TP(2) = y^1/y^{1*} = f^1\left(x^1\right)/f^0\left(x^1\right). \tag{19}$$

Thus, $TP(1)$ is one plus the percentage increase in output due to technical and managerial improvements (going from period 0 to period 1) evaluated at the period 0 input level x^0 and $TP(2)$ is one plus the percentage increase in output due to the new technology evaluated at the period 1 input level x^1.

It is also possible to define input-based measures of technical progress $TP(3)$ and $TP(4)$. First, define x^{0*} and x^{1*} as follows:

$$y^0 = f^1\left(x^{0*}\right) \text{ and } y^1 = f^0\left(x^{1*}\right). \tag{20}$$

Thus, x^{0*} is the input required to produce the period 0 output y^0 but by using the period 1 technology, and so x^{0*} will generally be less than x^0 (which is the amount of input required to produce the period 0 output using the period 0 technology). Similarly, x^{1*} is the amount of input required to produce the period 1 output y^1 but by using the period 0 technology, and x^{1*} will generally be larger than x^1 (because the period 0 technology will generally be less efficient than the period 1 technology). Now define the following two *input-based measures of technical progress*, $TP(3)$ and $TP(4)$[10]:

[9] $TP(1)$ and $TP(2)$ are the one input, one output special cases of Caves et al.'s (1982, p. 1402) output-based "productivity" indexes.

[10] $TP(3)$ and $TP(4)$ are the one input, one output special cases of Caves et al.'s (1982, p. 1407) input-based "productivity" indexes. However, in the present chapter, we regard these "productivity" indexes as measures of the shift in the production functions and hence as measures of technical progress.

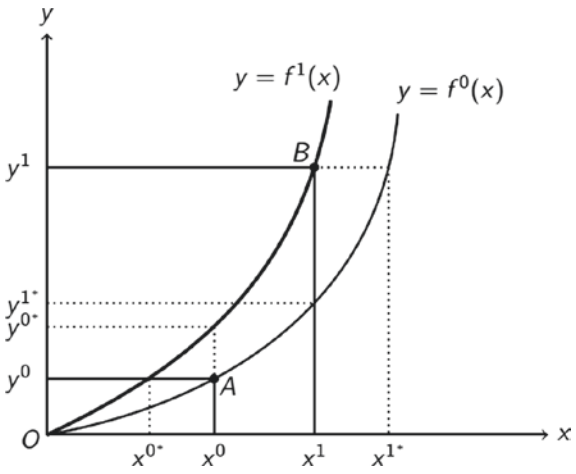

Fig. 1 Production-based measures of technical progress

$$TP(3) = x^0/x^{0*}; \tag{21}$$

$$TP(4) = x^{1*}/x^1. \tag{22}$$

The above four measures of *TP* can be illustrated with the aid of Fig. 1. The diagram shows that each of the *TP* measures can be different.

The lower curved line is the graph of the period 0 production function; that is, it is the set of points (x, y) such that $x \geq 0$ and $y = f^0(x)$. The higher curved line is the graph of the period 1 production function; that is, it is the set of points (x, y) such that $x \geq 0$ and $y = f^1(x)$. The observed data points are A, which has coordinates (x^0, y^0) and B, which has coordinates (x^1, y^1). Note that the absolute amounts of production function shift in the direction of the y-axis are $y^{0*} - y^0$(at point A) and $y^1 - y^{1*}$ (at point B). The absolute amounts of production function shift in the direction of the x-axis are $x^0 - x^{0*}$(at point A) and $x^{1*} - x^1$(at point B). We have chosen to measure *TP* in terms of the relative shifts, y^{0*}/y^0, y^1/y^{1*}, x^0/x^{0*} and x^{1*}/x^1 rather than the absolute shifts, $y^{0*} - y^0$, $y^1 - y^{1*}$, $x^0 - x^{0*}$ and $x^{1*} - x^1$ in order to obtain measures of shift that are invariant to changes in the units of measurement. Note that $MFPG = MFPG(2) = (y^1/x^1)/(y^0/x^0)$ is equal to the slope of the straight line OB divided by the slope of the straight line OA.

It turns out that there is a relationship between each of our technical progress measures, $TP(1)$, $TP(2)$, $TP(3)$, $TP(4)$, and $MFPG$. We have:

$$MFPG = TP(i)RS(i); \quad i = 1, 2, 3, 4 \tag{23}$$

where the four returns to scale measures $RS(i)$ are defined as follows:

$$RS(1) \equiv \left(y^1/x^1\right)/\left(y^{0*}/x^0\right); \tag{24}$$

$$RS(2) \equiv \left(y^{1*}/x^1\right)/\left(y^0/x^0\right); \tag{25}$$

$$RS(3) \equiv \left(y^1/x^1\right)/\left(y^0/x^{0*}\right); \tag{26}$$

$$RS(4) \equiv \left(y^1/x^{1*}\right)/\left(y^0/x^0\right). \tag{27}$$

The returns to scale measures $RS(1)$ and $RS(3)$ pertain to the period 1 production function f^1 while the measures $RS(2)$ and $RS(4)$ pertain to the period 0 production function f^0. To interpret each of these returns to scale measures geometrically, see Fig. 1. Each of these returns to scale measures is the ratio of two input-output coefficients, say y^j/x^j divided by y^k/x^k, the two points on the same production function, with $x^j > x^k$. Thus, if the returns to scale measure is greater than 1, then $y^j/x^j > y^k/x^k$ and we say that the production function exhibits increasing returns to scale between the two points. If $RS(i) = 1$, then the production function exhibits constant returns to scale between the two points and finally if $RS(i) < 1$, then the production function exhibits decreasing returns to scale between the two points.

The decompositions given by Eq. (23) tell us that *MFPG* is equal to the product of a technical progress term $TP(i)$ (this corresponds to a shift in the production function going from period 0 to period 1) and a returns to scale term $RS(i)$ (this corresponds to a movement along one of the production functions). In Fig. 1, definitions (18)–(22) and definitions (24)–(27) can be used to verify that each of the four decompositions of *MFPG* given by (23) corresponds to a different combination of shifts and movements along a production function that take us from point A to point B.

For firms in a regulated industry, returns to scale will generally be greater than one, since increasing returns to scale in production is often the reason for regulation in the first place. Thus, *MFPG* will exceed *TP* for growing firms in a regulated industry (provided that there are increasing returns to scale for that firm).

We note that the technical progress and returns to scale measures defined above cannot in general be calculated without a knowledge of the production functions that describe the technology for the two periods under consideration. However, in a one input, one output firm, the *MFPG* measures defined above can be calculated unambiguously provided that we know inputs used and outputs produced during the two periods.

Next, we shall generalize the above production function-based definitions of productivity and technical progress to cover the case of many outputs and many inputs.

3 Productivity Measurement in the Case of Many Outputs and Inputs

The approach taken in this section will be to replace the single output growth factor, y^1/y^0, by an *output quantity index* and to replace the single input growth factor, x^1/x^0, by an *input quantity index*. The approach outlined in this section is a practical one that is implemented by statistical agencies to calculate industry estimates of *MFP* growth.

Recall our first definition of productivity growth in the one output, one input case, $MFPG(1) \equiv \left(y^1/y^0\right)/\left(x^1/x^0\right)$, which was the output ratio divided by the input ratio between periods 0 and 1. In order to find a counterpart to this definition in the multiple output, multiple input case, we need only replace the output ratio by an output quantity index, $Q\left(p^0,p^1,y^0,y^1\right)$, and replace the input ratio by an input quantity index, $Q^*\left(w^0,w^1,x^0,x^1\right)$, where $p^t \equiv \left(p_1^t,\ldots,p_M^t\right)$ and $w^t \equiv \left(w_1^t,\ldots,w_N^t\right)$ are the period t output and input price vectors and $y^t \equiv \left(y_1^t,\ldots,y_M^t\right)$ and $x^t \equiv \left(x_1^t,\ldots,x_M^t\right)$ are the period t output and input quantity vectors for $t=0$, 1. Thus, an *output quantity index*, $Q\left(p^0,p^1,y^0,y^1\right)$, is defined to be a function of the output prices and quantities for the two periods under consideration. Similarly, an *input quantity index*, between periods 0 and 1, $Q^*\left(w^0,w^1,x^0,x^1\right)$, is simply a function of $4N$ variables, the input prices and quantities pertaining to the two periods under consideration.

Two of the most frequently used functional forms for quantity indexes are the Laspeyres (1871) and Paasche (1874) quantity indexes.[11] The *Laspeyres output quantity index* between periods 0 and 1 is defined as:

[11]Actually, Laspeyres and Paasche originally defined the price counterparts to the quantity indexes that we are defining here (see (41) and (42) below).

$$\begin{aligned} Q_L\left(p^0,p^1,y^0,y^1\right) &\equiv \sum_{m=1}^{M} p_m^0 y_m^1 / \sum_{m=1}^{M} p_m^0 y_m^0 \\ &= \sum_{m=1}^{M} \left(y_m^1/y_m^0\right) p_m^0 y_m^0 / \sum_{m=1}^{M} p_m^0 y_m^0 \\ &= \sum_{m=1}^{M} \left(y_m^1/y_m^0\right) s_m^0 \end{aligned} \tag{28}$$

where the *period t revenue share for output m* is defined as

$$s_m^t \equiv p_m^t y_m^t / \sum\nolimits_{k=1}^{M} p_k^t y_k^t; \quad m = 1,\ldots,M;\ t = 0,1. \tag{29}$$

Thus, the Laspeyres output quantity index is a base period revenue share weighted sum of the M individual quantity ratios, y_m^1/y_m^0.

The *Paasche output quantity index* between periods 0 and 1 is defined as:

$$\begin{aligned} Q_P\left(p^0,p^1,y^0,y^1\right) &\equiv \sum_{m=1}^{M} p_m^1 y_m^1 / \sum_{m=1}^{M} p_m^1 y_m^0 \\ &= \left(\sum_{m=1}^{M} p_m^1 y_m^0 / \sum_{m=1}^{M} p_m^1 y_m^1\right)^{-1} \\ &= \left[\sum_{m=1}^{M} \left(y_m^1/y_m^0\right)^{-1} p_m^1 y_m^1 / \sum_{m=1}^{M} p_m^1 y_m^1\right]^{-1} \\ &= \left[\sum_{m=1}^{M} \left(y_m^1/y_m^0\right)^{-1} s_m^1\right]^{-1}. \end{aligned} \tag{30}$$

Thus, the Paasche output quantity index is a current period revenue share weighted harmonic mean of the M individual quantity ratios, y_m^1/y_m^0.

In what follows, we shall concentrate on the problems involved in choosing a functional form for the output index Q; an analogous discussion applies to the choice of a functional form for the input index Q^*.

Another commonly used functional form for a quantity index is the Fisher (1922, p. 234) ideal quantity index Q_F which is equal to the square root of the product of the Laspeyres and Paasche quantity index defined by (28) and (30), i.e.:

$$Q_F\left(p^0,p^1,y^0,y^1\right) \equiv \left[Q_L\left(p^0,p^1,y^0,y^1\right)Q_P\left(p^0,p^1,y^0,y^1\right)\right]^{1/2}. \quad (31)$$

Another commonly used functional form for a quantity index is the Törnqvist (1936) quantity index Q_T. The natural logarithm of Q_T is defined to be the right-hand side of (32) below:

$$\ln Q_T\left(p^0,p^1,y^0,y^1\right) \equiv 1/2\sum_{m=1}^{M}\left(s_m^0+s_m^1\right)\ln\left(y_m^1/y_m^0\right) \quad (32)$$

where the revenue shares s_m^t are defined by (29) above. Note that the quantities y_m^t must all be positive in order for Q_T to be well defined.

The quantity index Q_T is also known as the *translog quantity index* (e.g. see Jorgenson and Nishimizu [1978] who introduced this terminology) because Diewert (1976, p. 120) related Q_T to a translog production function. This index is also known as the Divisia index since Jorgenson and Griliches (1967, 1972) used Q_T to provide a discrete time approximation to the continuous time Divisia index.[12]

The four quantity indexes Q_L, Q_P, Q_F and Q_T, defined by (28), (30), (31) and (32) respectively, all have a common property: if the number of outputs M equals one, then each of these quantity indexes reduces to the output ratio, y_1^1/y_1^0. Thus, it can be seen that the use of quantity indexes for outputs and inputs can be used to generalize our one output, one input measure of productivity change, *MFPG*(1), discussed in the previous section. More formally, let us define the direct quantity index measure of productivity growth *MFPG*(5) in the general multiple output, multiple input case as follows:

$$MFPG(5) \equiv Q\left(p^0,p^1,y^0,y^1\right)/Q^*\left(w^0,w^1,x^0,x^1\right) \quad (33)$$

where Q is the output quantity index and Q^* is the input quantity index. If the number of outputs equals one and the number of inputs equals one, if Q equals one of Q_L, Q_P, Q_F or Q_T, and if Q^* equals one of Q_L^*, Q_P^*, Q_F^* or Q_T^*, then $MFPG(5) = MFPG(1)$. Thus, the approach to productivity measurement outlined in this section reduces to the approach outlined in the previous section if there is only one input and only one output.

[12]Unfortunately, there are many discrete time approximations to the Divisia index including the Paasche and Laspeyres quantity indexes (see Frisch 1936; Diewert 1980).

In the general multiple output, multiple input case, we still have to address a problem: Which functional forms for the output index Q and the input index Q^* should we choose? We shall return to this functional form problem shortly.

We turn now to an index number measure of productivity that generalizes the deflated revenues divided by deflated costs productivity measure $MFPG(3)$ that was defined earlier by (7) in the previous section.

Denote period t revenue by R^t and period t cost by C^t. We have:

$$R^t \equiv \sum_{m=1}^{M} p_m^t y_m^t; \; C^t \equiv \sum_{n=1}^{N} w_n^t x_n^t; \quad t = 0, 1. \tag{34}$$

The multiple output analogue to the output price ratio which occurred in formula (34) in the previous section is the o*utput price index*, $P(p^0, p^1, y^0, y^1)$, which is a function of $4M$ variables, the output prices and quantities that pertain to the two periods under consideration. The multiple input analogue to the input price ratio which occurred in the previous section is the *input price index*, $P^*(w^0, w^1, x^0, x^1)$, which is a function of $4N$ variables, the input prices and quantities that pertain to the two periods under consideration.

Using the output price index P as a deflator for the revenue ratio R^1/R^0 between periods 0 and 1 and using the input price index P^*as a deflator for the cost ratio C^1/C^0 between the two periods leads to the following definition of the productivity growth of the production unit going from period 0 to 1:

$$MFPG(6) \equiv \left[\left(R^1/R^0\right)/P\left(p^0, p^1, y^0, y^1\right)\right]/\left[\left(C^1/C^0\right)/P^*\left(w^0, w^1, x^0, x^1\right)\right]. \tag{35}$$

Note that (35) is a generalization to multiple inputs and outputs of our earlier productivity change measure $MFPG(3)$ defined in the previous section.

Suppose that the output quantity index $Q(p^0, p^1, y^0, y^1)$ which appeared in definition (33) matches up with the output price index $P(p^0, p^1, y^0, y^1)$ which appears in definition (35) in the sense that the product of the price and quantity index equals the revenue ratio for the two periods under consideration so that we have:

$$R^1/R^0 = P\left(p^0, p^1, y^0, y^1\right) Q\left(p^0, p^1, y^0, y^1\right). \tag{36}$$

Suppose further that the input quantity index $Q^*(w^0, w^1, x^0, x^1)$ which appeared in definition (33) matches up with the input price index $P^*(w^0, w^1,$

x^0, x^1) which appears in definition (35) in the sense that the product of the price and quantity index equals the cost ratio for the two periods under consideration so that we have:

$$C^1/C^0 = P^*\left(w^0, w^1, x^0, x^1\right) Q^*\left(w^0, w^1, x^0, x^1\right). \tag{37}$$

Now substitute (36) and (37) into (35) and we find that:

$$MFPG(5) = MFPG(6). \tag{38}$$

Thus if the two pairs of price and quantity indexes satisfy the relations (36) and (37), we find that both of the productivity measures introduced in this section, *MFPG*(5) defined by (33) and *MFPG*(6) defined by (35) are equal to each other.

Recall that in the previous section, we defined the period t markup, m^t, for the production unit by $1+m^t = R^t/C^t$ for $t=0,1$. Using these definitions of the markup in each period again, it can be seen that we can rewrite *MFPG*(6) as follows:

$$\begin{aligned} MFPG(6) &= \left[\left(R^1/R^0\right)/P\left(p^0,p^1,y^0,y^1\right)\right]/\left[\left(C^1/C^0\right)/P^*\left(w^0,w^1,x^0,x^1\right)\right] \\ &= \left[\left(R^1/R^0\right)/\left(C^1/C^0\right)\right]\left[P^*\left(w^0,w^1,x^0,x^1\right)/P\left(p^0,p^1,y^0,y^1\right)\right] \\ &= \left[\left(1+m^1\right)/\left(1+m^0\right)\right]\left[P^*\left(w^0,w^1,x^0,x^1\right)/P\left(p^0,p^1,y^0,y^1\right)\right] \\ &= MFPG(7). \end{aligned} \tag{39}$$

The above definition says that *MFPG*(7) is equal to the margin growth rate times the input price index divided by the output price index. Defining *profitability* as R^t/C^t for $t=0,1$, we can see from the second line of (39) that we have productivity growth equal to the growth in profitability times the relative growth of input prices to output prices. Reorganizing, we get that profitability growth equals productivity growth times the relative growth of output prices to input prices. This highlights the role of productivity as a key determinant of profitability (see, e.g., Balk [2003, p. 2] for more on this).

Note that *MFPG*(7) is an exact analogue to our earlier one output, one input *MFP* growth measure *MFPG*(4) defined by (12) in Sect. 1. Equations (38) and (39) show that this "new" measure of *MFP* growth is equal to the previous measure *MFPG*(5), which was the ratio of the output quantity index to the input quantity index, and to *MFPG*(6), which was equal to the revenue growth rate deflated by the output price index divided

by the cost growth rate deflated by the input price index.[13] Thus, we have obtained multiple output, multiple input counterparts to the following Sect. 1 equalities:

$$MFPG(1) = MFPG(3) = MFPG(4). \tag{40}$$

There remains the problem of choosing a functional form for the output price index P and the input price index P^*. The same four index number formulae that were used for quantity indexes, (28), (30), (31) and (32), can also be used for price indexes, except that the role of prices and quantities are interchanged. Thus, define the Laspeyres price index P_L, the Paasche price index P_P, the Fisher price index P_F and the translog price index P_T by (41), (42), (43) and (44), respectively:

$$P_L\left(p^0,p^1,y^0,y^1\right) \equiv Q_L\left(y^0,y^1,p^0,p^1\right); \tag{41}$$

$$P_P\left(p^0,p^1,y^0,y^1\right) \equiv Q_P\left(y^0,y^1,p^0,p^1\right); \tag{42}$$

$$P_F\left(p^0,p^1,y^0,y^1\right) \equiv Q_F\left(y^0,y^1,p^0,p^1\right); \tag{43}$$

$$P_T\left(p^0,p^1,y^0,y^1\right) \equiv Q_T\left(y^0,y^1,p^0,p^1\right). \tag{44}$$

The Laspeyres, Paasche, Fisher and Translog input price indexes, $P_L^*\left(w^0,w^1,x^0,x^1\right)$, $P_P^*\left(w^0,w^1,x^0,x^1\right)$, $P_F^*\left(w^0,w^1,x^0,x^1\right)$, and $P_T^*\left(w^0,w^1,x^0,x^1\right)$, respectively, may be defined in an analogous manner.

If $M=1$, so that there is only one output, then it can be verified that the output price indexes defined by (41)–(44) all collapse down to the output price ratio, p_1^1/p_1^0. Similarly, if $N=1$, so that there is only one input, then P_L^*, P_P^*, P_F^* and P_T^* all collapse down to the input price ratio, w_1^1/w_1^0. Thus, the use of the Laspeyres, Paasche, Fisher or translog price indexes in (35) or (39) leads to the following equalities in the $M=1$, $N=1$:

$$MFPG(6) = MFPG(7) = MFPG(1). \tag{45}$$

Thus, our new definitions of productivity change defined by (33), (35) or (39) are generalizations to the case of many outputs and inputs of our earlier

[13] We require that (36) and (37) hold in order to obtain these equalities.

one output, one input measure of productivity change defined by (3) in the previous section.

Returning to the general case of many outputs and many inputs, it can be seen that different choices of the output price index P and the input price index P^* will generate different productivity change measures $MFPG(6)$ defined by (35). Similarly, different choices of the output quantity index Q and the input quantity index Q^* will generate different productivity change measures $MFPG(5)$ defined by (33).

However, the degree of arbitrariness in the formulae (33) and (35) is not quite as large as it might seem at first glance. It turns out that the two families of productivity measures are related, because the deflated revenue ratio which occurs in the numerator of the right-hand side of (35), $\left(R^1/R^0\right)P\left(p^0,p^1,y^0,y^1\right)$, can be interpreted as an *implicit quantity index of outputs*, and the denominator in (35), $\left(C^1/C^0\right)P^*\left(w^0,w^1,w^0,w^1\right)$, can be interpreted as an *implicit quantity index of inputs*.

From Sect. 2, it was evident that the total factor productivity growth measures that were defined there measure the combined effects of technological progress, movements towards the production frontier and increasing (or decreasing) returns to scale. The *MFP* growth measures defined in this section also measure the combined effects of these three factors. When we allow for the possibility of increasing returns to scale in production, it turns out to be very difficult to estimate separately the effects of increasing returns to scale from technical progress. In general, in order to perform this separation, it is necessary to have panel data or to perform an econometric study on time series data.[14] Econometric approaches are, in general, not practical for a statistical agency. And usually, statistical agencies do not have usable panel data on hand in order to undertake nonparametric studies of relative efficiency. Hence, we will not cover these econometric approaches and applications of nonparametric methods utilizing cross-sectional data in this brief survey of how to measure *MFP* growth.[15]

In the next two sections, we turn to an assessment of the alternative index number formulae introduced in this section. This will help explain the

[14]See Basu and Fernald (1997, 2002), Lawrence and Diewert (2006) and Diewert and Fox (2008) for econometric methods that can estimate the separate contributions of technical progress and returns to scale in the time series context. Their work draws on the earlier work of Nakajima et al. (1998, 2002).

[15]There is a huge literature on the nonparametric approach to measuring productivity and efficiency (see, e.g., Farrell (1957), Afriat (1972), Charnes et al. (1978), Diewert and Parkan (1983), Varian (1984), Färe (1988), Balk (1998, 2003), Diewert and Nakamura (1999), Diewert and Mendoza (2007) and Diewert and Fox (2014, 2017, 2018a).

properties of index number formulae used by NSOs, and why some formulae are favoured over others.

4 The Test Approach to Index Number Theory

First, we introduce another index number formula to be assessed. It can be shown that $(R^1/R^0)/P_T(p^0,p^1,y^0,y^1)$ is *not* equal to the Törnqvist quantity index, Q_T. Hence, we simply define the *implicit Törnqvist quantity index*, Q_{IT}, as follows:

$$Q_{IT}\left(p^0,p^1,y^0,y^1\right) \equiv \left(R^1/R^0\right)/P_T\left(p^0,p^1,y^0,y^1\right). \tag{46}$$

The five quantity indexes, Q_L, Q_P, Q_F, Q_T and Q_{IT}, are the five functional forms for quantity indexes that are used most frequently in applied economics. The question now arises: Which of these five formulae should we use in the multiple output, multiple input definition of *MFP* growth, *MFPG*(5) defined by (35)?

Using the results from Diewert (1976), it can be shown that from the perspective of the economic approach to index number theory, Q_F, Q_T and Q_{IT} are clearly preferred to the Paasche and Laspeyres quantity indexes, Q_P and Q_L. Again, from the perspective of the economic approach to index number theory, P_F, P_T and P_{IT} are clearly preferred to the Paasche and Laspeyres price indexes, P_P and P_L. The economic approach provides equal justifications for Q_F, Q_T and Q_{IT} or for P_F, P_T and P_{IT}. Hence, any of these indexes would be equally good from the economic perspective.[16] We will pursue the economic approach in more detail in the following section.

Another major approach to index number theory is the *test or axiomatic approach* to index number theory. This approach to the determination of the functional form for P and Q works as follows: researchers suggest various mathematical properties that P or Q should satisfy based on a priori reasoning—these properties are called "tests" or "axioms"—and then mathematical reasoning is applied to determine: (i) whether the a priori tests are mutually consistent and (ii) whether the a priori tests uniquely determine

[16]Diewert (1978) showed that for normal time series data, all of these indexes give much the same answer since they approximate each other to the second order around an equal price and quantity point.

the functional form for P or Q. The main contributors to the test or axiomatic approach were Walsh (1901, 1921a, b), Fisher (1911, 1922), Frisch (1936), Eichhorn (1978), Eichhorn and Voeller (1976) and Funke and Voeller (1978, 1979).[17]

We will not cover the test approach in great detail in this chapter but we will present some material on this important approach to index number theory.

One fundamental test that the price and quantity index should jointly satisfy is the test (36) above; that is, the product of the output price and quantity indexes between periods 0 and 1 should equal the revenue or value ratio between the two periods, $R_1/R_0 = \sum_{m=1}^{M} p_m^1 y_m^1 / \sum_{m=1}^{M} p_m^0 y_m^0$. This test was called the *product test* by Frisch (1930, p. 399), but it was first formulated by Irving Fisher (1911, p. 388).

If we accept the validity of the product test (and virtually all researchers do accept its validity), then P and Q cannot be determined independently. For example, if the functional form for the price index P is given, then (36) determines the functional form for the quantity index Q.

Thus, in what follows, we focus on the determination of the functional form for the price index P. Once P has been determined, Q will be determined residually by (36).

We list a few examples of tests that have been proposed for price indexes.

The *Identity* or *Constant Prices Test*, originally proposed by Laspeyres (1871, p. 308) and also by Walsh (1901, p. 308), and Eichhorn and Voeller (1976, p. 24) is the following test:

$$P\left(p, p, y^0, y^1\right) = 1; \tag{47}$$

i.e., if $p^0 = p^1 \equiv p$, so that for each commodity, prices are equal in the two periods being compared, then the price index is equal to 1 no matter what the quantities are in period 0 and 1, y^0 and y^1 respectively.

The *Constant Basket Test* or the *Constant Quantities Test*, proposed by many researchers including Walsh (1901, p. 540), is the following test:

$$P\left(p^0, p^1, y, y\right) = \sum_{m=1}^{M} p_m^1 y_m / \sum_{m=1}^{M} p_m^0 y_m; \tag{48}$$

[17]For more recent contributions and surveys, see Diewert (1992, 1993, 1997, 2008) and Balk (1995, 2008).

i.e., if quantities are constant over the two periods 0 and 1 so that $y^0 = y^1 \equiv y$, then the level of prices in period 1 compared to period 0 is the value of the constant basket of quantities evaluated at the period 1 prices, $\sum_{m=1}^{M} p_m^1 y_m$, divided by the value of the basket evaluated at the period 0 prices, $\sum_{m=1}^{M} p_m^0 y_m$.

The *Proportionality in Period t Prices Test*, proposed by Walsh (1901, p. 385) and Eichhorn and Voeller (1976, p. 24), is the following test:

$$P(p^0, \lambda p^1, y^0, y^1) = \lambda P\left(p^0, p^1, y^0, y^1\right) \quad \text{for all } \lambda > 0; \tag{49}$$

i.e., if each price in period 1 is multiplied by the positive constant λ, then the level of prices in period 1 relative to the level of prices in period 0 increases by the same positive constant λ.

Our final example of a price index test is the *Time Reversal Test*, which was first informally proposed by Pierson (1896, p. 128) and more formally by Walsh (1901, p. 368; 1921b, p. 541) and Fisher (1922, p. 64):

$$P\left(p^1, p^0, y^1, y^0\right) = 1/P\left(p^0, p^1, y^0, y^1\right); \tag{50}$$

i.e., if the prices and quantities for periods 0 and 1 are interchanged, then the resulting price index is the reciprocal of the original price index.

The four tests (47)–(50) will suffice to give a flavour of the test approach to index number theory. For a much more extensive list of twenty or so tests (see Diewert (1992)).

There are five leading functional forms for the output price index P that are most frequently used in empirical work: (i) the Laspeyres price index P_L, (ii) the Paasche price index P_P, (iii) the Fisher price index P_F, (iv) the Törnqvist price index P_T defined by (44) and (v) the implicit Törnqvist price index P_{IT} defined by:

$$P_{IT}\left(p^0, p^1, y^0, y^1\right) \equiv \left[\sum_{m=1}^{M} p_m^1 y_m^1 / \sum_{m=1}^{M} p_m^0 y_m^0\right] / Q_T\left(p^0, p^1, y^0, y^1\right) \tag{51}$$

where the Törnqvist quantity index Q_T is defined by (32). The Fisher index satisfies the four tests (47)–(50), but P_L fails (50), P_P fails (50), P_T fails (48) and P_{IT} fails (47).

When more extensive lists of tests are compiled, the Fisher ideal price index P_F continues to satisfy more tests than other leading candidates (see Diewert 1976, p. 131; 1992). In fact, the Fisher price index satisfies all twenty tests utilized by Diewert (1992). Moreover, satisfactory axiomatic

characterizations of P_F have been obtained (see Funke and Voeller 1978, p. 180; 1979; Diewert 1992). Thus, from the viewpoint of the test approach to index number theory, the Fisher quantity index Q_F defined by (31) and the corresponding Fisher price index P_F defined by (43) seem to be the best choices. It should also be noted that P_F and Q_F satisfy the product test in (36). Hence, if the Fisher indexes are used in the productivity measures defined by (33) or (35), then both of these productivity measures will coincide; that is, if we use Fisher price and quantity indexes for P and Q and P^* and Q^* wherever they occur in (33), (35) or (39), we obtain the following equality:

$$MFPG_F(5) = MFPG_F(6) = MFPG_F(7) \tag{52}$$

where we have added a subscript F to the three productivity measures to indicate that Fisher indexes are being used. *Thus, an added benefit of using Fisher price and quantity indexes is that three conceptually distinct* (but equally attractive) *productivity change measures become identical*.

While the Törnqvist index fails nine of twenty tests of Diewert (1992), it passes the time reversal test, which is regarded as an important property. Also, it usually approximates the Fisher index closely in empirical applications, so we can regard it as satisfying all twenty tests to a high degree of approximation. The Laspeyres and Paasche indexes fail only three ("reversal") tests, but the failure to satisfy the time reversal test is regarded as serious. Hence, from the test approach to index numbers, the Fisher and Törnqvist indexes are preferred.

In the next section, we look at an index number method for estimating *MFP* growth in the time series context that draws on the economic approach to the measurement of *MFP* growth and the theory of exact index numbers.

5 The Exact Index Number Approach to Productivity Measurement

The test approach to index number choice discussed in Sect. 4 related to the mathematical properties of the index formulae. There was no direct connection with economic theory. However, such a connection can be made, as will be shown in this section. This "economic" or "exact" approach to index number choice has been influential in guiding index number choice by NSOs and is a reason why the USA switched to using a Fisher index

formula for calculating gross domestic product (GDP) in the mid-1990s. It is also a reason why it is common practice to use Törnqvist in constructing industry-level *MFP* estimates.[18]

Konüs (1939) introduced the idea of a *true cost of living index*, which is a ratio of cost functions where utility is held constant. The corresponding concept in the production context is that the true price index is the ratio of revenue functions $R(p^t,y)/R(p^{t-1},y)$, where y is a reference output level. For a choice of functional form for the revenue functions, this unobserved theoretical true price index can be exactly calculated. In this case, we say that there is an "exact" relationship between the functional form and an index number formula. For example, it can be shown that for a linearly homogeneous quadratic unit revenue function, assuming optimizing behaviour (so that Hotelling's Lemma can be used), the true price index exactly equals the Fisher price index.[19]

The justification for the Törnqvist index can be argued to be stronger than for the Fisher index from this approach, as the assumption of linear homogeneity is not required to establish its exact relationship with the translog functional form. Both translog and quadratic functional forms have the property of "flexibility"; Diewert (1974) defined a *flexible functional form* as one that provides a second-order approximation to a twice continuously differentiable function at a point. Many popular functional forms in economics (e.g. Cobb-Douglas and CES) do not have this rather minimal property. An index number which is exact for a flexible functional form was defined by Diewert (1976) as being *superlative*. Thus, Fisher and Törnqvist indexes are superlative indexes.

Laspeyres and Paasche indexes are not superlative. They are exact for a linear unit cost function (Konüs and Byushgens 1926), which is dual to a (zero substitution) Leontief production function. Thus, these indexes are regarded as quite restrictive from the economic approach to index numbers.

In this section, we appeal to the exact index number approach to develop our approach to measuring *MFP* growth when there are many outputs and many inputs. We describe the exact index number approach to the measurement of technical change and productivity growth that was initially developed by Diewert and Morrison (1986) and Kohli (1990). This theory

[18]Both the U.S. Bureau of Labor Statistics and the Australian Bureau of Statistics use the Törnqvist formula for constructing MFP estimates (see BLS [n.d.] and Moulton [2018] for the U.S. and ABS [2015a, 2018a] for Australia).

[19]Drawing on results from Byushgens (1925), Konüs and Byushgens (1926) and Diewert (1976, pp. 133–134) obtained this result in the consumer context.

is adapted into a method for measuring the growth in the real income generated by a production unit with a decomposition of this growth in real income into components that reflect:

- technical progress;
- changes in the prices of outputs; and
- growth of primary inputs.

This methodology can provide measures of how changes in the prices of imports and exports can affect real income growth.

We assume that there is a period t market sector technology set S^t that exhibits constant returns to scale. The components of net output are the usual components of GDP, namely $C+G+I+X-M$ (household and government consumption, investment, exports minus imports). Later we will also subtract depreciation and revaluation terms from GDP in order to obtain net domestic product, which is closer to an income concept. For now, we interpret the net output vector for period t, y^t, as the net output components of market sector of the economy. The corresponding market sector primary input vector for period t is denoted by x^t. The components of x^t consist of different types of labour services supplied to the market sector by households and the various types of capital services used by the market sector. The corresponding vectors of period t net output prices is denoted by P^t and the corresponding vector of period t primary input prices is denoted by W^t. In period t, we assume that there is a feasible set of output vectors y that can be produced by the market sector if the vector of primary inputs x is utilized by the market sector of the economy; denote this period t production possibilities set by S^t. We assume that S^t is a closed convex cone that exhibits a free disposal property.[20]

[20]For more explanation of the meaning of these properties, Diewert (1973, 1974, p. 134) or Woodland (1982) or Kohli (1978, 1991). The assumption that S^t is a cone means that the technology is subject to constant returns to scale. This is an important assumption since it implies that the value of outputs should equal the value of inputs in equilibrium. In empirical work, this property can be imposed upon the data by using an ex post rate of return in the user costs of capital, which forces the value of inputs to equal the value of outputs for each period. The function g^t is known as the *GDP function* or the *gross national product function* in the international trade literature (see Kohli 1978, 1991, 2004a, b; Woodland 1982; Feenstra 2004, p. 76). It was introduced into the economics literature by Samuelson (1953). Alternative terms for this function include: (i) the *gross profit function* (see Gorman 1968); (ii) the *restricted profit function* (see Lau 1976; McFadden 1978); and (iii) the *variable profit function* (see Diewert (1973, 1974).

Given a vector of output prices P and a vector of available primary inputs x, we define *the period t market sector GDP function*, $g^t(P,x)$, as follows[21]:

$$g^t(P,x) \equiv \max_y \{Py : (y,x) \text{ belongs to } S^t\}; \quad t = 1,2,\ldots. \tag{53}$$

Thus, market sector GDP depends on t (which represents the period t technology set S^t), on the vector of output prices P that the market sector faces and on x, the vector of primary inputs that is available to the market sector.

If P^t is the period t output price vector and x^t is the vector of inputs used by the market sector during period t and assuming that actual outputs equal the theoretical market sector outputs given by the solution to Eq. (53), then the period t vector of market sector outputs y^t will be equal to the vector of first-order partial derivatives of $g^t(P^t,x^t)$ with respect to the components of P; that is, we will have the following equations for each period t[22]:

$$y^t = \partial_P g^t\left(P^t,x^t\right); \quad t = 1,2,\ldots. \tag{54}$$

Thus, assuming profit maximization, the period t market sector (net) supply vector y^t can be obtained by differentiating the period t market sector GDP function with respect to the components of the period t output price vector P^t.

Assuming that actual primary inputs equal the theoretical market sector inputs that minimize the cost of producing a given amount of GDP, then the period t vector of input prices W^t will be equal to the vector of first-order partial derivatives of $g^t(P^t,x^t)$ with respect to the components of x; that is, we will have the following equations for each period t[23]:

$$W^t = \partial_x g^t\left(P^t,x^t\right); \quad t = 1,2,\ldots. \tag{55}$$

Thus, assuming cost minimization, the period t market sector input prices W^t paid to primary inputs can be obtained by differentiating the period t

[21] The function $g^t(P, x)$ will be linearly homogeneous and convex in the components of P and linearly homogeneous and concave in the components of x (see Diewert 1973, 1974, p. 136). Notation: $Py \equiv \sum_{m=1}^{M} P_m y_m$.

[22] These relationships are due to Hotelling (1932, p. 594). Note that $\nabla_p g^t\left(P^t,x^t\right) \equiv \left[\partial g^t\left(P^t,x^t\right)/\partial P_1,\ldots,\partial g^t\left(P^t,x^t\right)/\partial P_M\right]$.

[23] These relationships are due to Samuelson (1953) and Diewert (1974, p. 140). Note that $\nabla_x g^t\left(P^t,x^t\right) \equiv \left[\partial g^t\left(P^t,x^t\right)/\partial x_1,\ldots,\partial g^t\left(P^t,x^t\right)/\partial x_N\right]$.

market sector GDP function with respect to the components of the period t input quantity vector x^t.

The assumptions of price-taking behaviour in relating quantities to prices, i.e., the assumption of pure competition, will be maintained in the remainder of this chapter. The fascinating violations of this assumption are analysed in Chapters 13 and 15 of this Handbook.

The constant returns to scale assumption on the technology sets S^t implies that the value of outputs will equal the value of inputs in period t; that is, we have the following relationships:

$$g^t\left(P^t, x^t\right) = P^t \cdot y^t = W^t \cdot x^t; \quad t = 1, 2, \ldots. \tag{56}$$

This says that nominal GDP constructed using the production approach (value of outputs) should equal GDP constructed using the income approach (payments to the factors of production). NSOs typically aim to ensure that this is the case. Whether or not the assumption of constant returns to scale is desirable could be questioned, as it forces the value of output to equal the value of input, but here we simply note that it is standard NSO practice to do so.[24]

Our focus is on the income generated by the market sector or more precisely, on *the real income generated by the market sector*. However, since market sector net output is distributed to the factors of production used by the market sector, nominal market sector GDP will be equal to nominal market sector income, as in (56). As an approximate welfare measure that can be associated with market sector production,[25] we will choose to measure the *real income generated by the market sector in period t*, ρ^t, in terms of the

[24]At issue is whether, in calculating costs, we should use a endogenous balancing rate of return in the user cost of capital formula or an exogenous one (see Diewert and Fox [2018b] for (much) more on the calculation of user costs). Both approaches are used. For example, the ABS and Statistics New Zealand use a mixture of endogenous and exogenous rates, through placing a floor the rate of return as CPI plus 4% (see ABS 2015a). The advantage of the balancing rate approach is that we do not have to introduce a pure profits cell into the production accounts (which is problematic when it comes to deflating this nominal cell in the "real" accounts). Do we use a Consumer Price Index to deflate this balancing pure profits (or losses) item or what is the alternative? We do have to force *ex post* balance between the nominal value of output and the nominal value of input plus net pure profits? There are also unresolved issues when we have increasing returns to scale (or decreasing costs due to large fixed costs) (see, e.g., Diewert and Fox 2008).

[25]Since some of the primary inputs used by the market sector can be owned by foreigners, our measure of *domestic* welfare generated by the market production sector is only an approximate one. Moreover, our suggested welfare measure is not sensitive to the distribution of the income that is generated by the market sector.

number of consumption bundles that the nominal income could purchase in period t; that is, define ρ^t as follows:

$$\rho^t \equiv W^t \cdot x^t / P_C^t = w^t \cdot x = p^t \cdot y^t = g^t\left(p^t, x^t\right); \quad t = 0, 1, 2, \ldots \tag{57}$$

where $P_C^t > 0$ is the *period t consumption expenditures deflator* and the market sector period t *real output price* p^t and *real input price* w^t vectors are defined as the corresponding nominal price vectors deflated by the consumption expenditures price index; that is, we have the following definitions[26]:

$$p^t \equiv P^t / P_C^t; w^t \equiv W^t / P_C^t; \quad t = 0, 1, 2, \ldots. \tag{58}$$

The first and last equality in (57) imply that period t real income, ρ^t, is equal to the period t GDP function, evaluated at the period t real output price vector p^t and the period t input vector x^t, $g^t(p^t,x^t)$. Thus, *the growth in real income over time can be explained by three main factors: Technical Progress or Total Factor Productivity growth,*[27] *growth in real output prices and the growth of primary inputs*. We will shortly give formal definitions for these three growth factors.

Using the linear homogeneity properties of the GDP functions $g^t(P,x)$ in P and x separately, we can show that the following counterparts to the relations (54) and (55) hold using the deflated prices p and w[28]:

$$y^t = \nabla_p g^t\left(p^t, x^t\right); \quad t = 0, 1, 2, \ldots \tag{59}$$

$$w^t = \nabla_x g^t\left(p^t, x^t\right); \quad t = 0, 1, 2, \ldots. \tag{60}$$

[26]This approach is similar to the approach advocated by Kohli (2004b, 92), except he essentially deflated nominal GDP by the domestic expenditures deflator rather than just the domestic (household) expenditures deflator; i.e., he deflated by the deflator for $C+G+I$, whereas we suggest deflating by the deflator for C. Another difference in his approach compared to the present approach is that we restrict our analysis to the market sector GDP, whereas Kohli deflates all of GDP (probably due to data limitations). Our treatment of the balance of trade surplus or deficit is also different.

[27]Technical progress and MFP (and hence TFP) are synonymous here due to the assumption of constant returns to scale.

[28]If producers in the market sector of the economy are solving the profit maximization problem that is associated with $g^t(P, x)$, which uses the original output prices P, then they will also solve the profit maximization problem that uses the normalized output prices $p \equiv P/P_C$; i.e., they will also solve the problem defined by $g^t(p, x)$.

Now we are ready to define a family of period t productivity growth factors or technical progress shift factors $\tau(p,x,t)$[29]:

$$\tau(p,x,t) \equiv g^t(p,x)/g^{t-1}(p,x); \quad t = 1,2,\ldots. \tag{61}$$

Thus, $\tau(p,x,t)$ measures the proportional change in the real income produced by the market sector at the reference real output prices p and reference input quantities used by the market sector x where the numerator in (61) uses the period t technology and the denominator in (61) uses the period $t-1$ technology. Thus, each choice of reference vectors p and x will generate a possibly different measure of the shift in technology going from period $t-1$ to period t. Note that we are using the chain system to measure the shift in technology.

It is natural to choose special reference vectors for the measure of technical progress defined by (61): a *Laspeyres type measure* τ_L^t that chooses the period $t-1$ reference vectors p^{t-1} and x^{t-1} and a *Paasche type measure* τ_P^t that chooses the period t reference vectors p^t and x^t:

$$\begin{aligned} \tau_L^t &\equiv \tau\left(p^{t-1},x^{t-1},t\right) \\ &= g^t\left(p^{t-1},x^{t-1}\right)/g^{t-1}\left(p^{t-1},x^{t-1}\right); \quad t = 1,2,\ldots; \end{aligned} \tag{62}$$

$$\tau_P^t \equiv \tau\left(p^t,x^t,t\right) = g^t\left(p^t,x^t\right)/g^{t-1}\left(p^t,x^t\right); \quad t = 1,2,\ldots. \tag{63}$$

Since both measures of technical progress are equally valid, it is natural to average them to obtain an overall measure of technical change. If we want to treat the two measures in a symmetric manner and we want the measure to satisfy the time reversal property from the index number theory in Sect. 2, then the geometric mean will be the best simple average to take.[30] Thus, we define the geometric mean of (62) and (63) as follows[31]:

$$\tau^t \equiv \left(\tau_L^t \tau_P^t\right)^{1/2}; \quad t = 1,2,\ldots. \tag{64}$$

[29]This measure of technical progress is due to Diewert (1983, p. 1063) and Diewert and Morrison (1986, p. 662). Salter (1960) introduced the analogous measure for cost functions.

[30]See the discussion in Diewert (1997) on choosing the "best" symmetric average of Laspeyres and Paasche indexes that will lead to the satisfaction of the time reversal test by the resulting average index.

[31]The specific theoretical productivity change indexes defined by (62)–(64) were first defined by Diewert and Morrison (1986, pp. 662–663). See Diewert (1993) for properties of symmetric means.

At this point, it is not clear how we will obtain empirical estimates for the theoretical productivity growth factors defined by (62)–(64). One obvious way would be to assume a functional form for the GDP function $g^t(p,x)$, collect data on output and input prices and quantities for the market sector for a number of years (and for the consumption expenditures deflator), add error terms to Eqs. (59) and (60) and use econometric techniques to estimate the unknown parameters in the assumed functional form. However, econometric techniques are generally not completely straightforward: different econometricians will make different stochastic specifications and will choose different functional forms.[32] Moreover, as the number of outputs and inputs grows, it will be impossible to estimate a flexible functional form. Thus, we will suggest methods for estimating measures like (64) that are based on exact index number techniques.

We turn now to the problem of defining theoretical indexes for the effects on real income due to changes in real output prices. Define a family of *period t real output price growth factors* $\alpha\left(p^{t-1},p^t,x,s\right)$[33]:

$$\alpha\left(p^{t-1},p^t,x,s\right) \equiv g^s\left(p^t,x\right)/g^s\left(p^{t-1},x\right); \quad s=1,2,\ldots. \tag{65}$$

Thus, $\alpha\left(p^{t-1},p^t,x,s\right)$ measures the proportional change in the real income produced by the market sector that is induced by the change in real output prices going from period $t-1$ to t, using the technology that is available during period s and using the reference input quantities x. Thus, each choice of the reference technology s and the reference input vector x will generate a possibly different measure of the effect on real income of a change in real output prices going from period $t-1$ to period t.

Again, it is natural to choose special reference vectors for the measures defined by (65): a *Laspeyres type measure* α_L^t that chooses the period $t-1$ reference technology and reference input vector x^{t-1} and a *Paasche type measure*

[32]"The estimation of GDP functions…can be controversial, however, since it raises issues such as estimation technique and stochastic specification. … We therefore prefer to opt for a more straightforward index number approach" (Kohli 2004a, p. 344).

[33]This measure of real output price change was essentially defined by Fisher and Shell (1972, pp. 56–58), Samuelson and Swamy (1974, pp. 588–592), Archibald (1977, pp. 60–61), Diewert (1980, pp. 460–461; 1983, p. 1055) and Balk (1998, pp. 83–89). Readers who are familiar with the theory of the true cost of living index will note that the real output price index defined by (65) is analogous to the Konüs (1939) *true cost of living index* which is a ratio of cost functions, say $C(u, p^t)/C(u, p^{t-1})$ where u is a reference utility level: g^t replaces C and the reference utility level u is replaced by the vector of reference variables x.

α_P^t that chooses the period t reference technology and reference input vector x^t:

$$\begin{aligned}\alpha_L^t &= \alpha\left(p^{t-1}, p^t, x^{t-1}, t-1\right) \\ &= g^{t-1}\left(p^t, x^{t-1}\right)/g^{t-1}\left(p^{t-1}, x^{t-1}\right); \quad t = 1, 2, \ldots;\end{aligned} \tag{66}$$

$$\alpha_P^t = \alpha\left(p^{t-1}, p^t, x^t, t\right) = g^t\left(p^t, x^t\right)/g^t\left(p^{t-1}, x^t\right); \quad t = 1, 2, \ldots. \tag{67}$$

Since both measures of real output price change are equally valid, it is natural to average them to obtain an overall measure of the effects on real income of the change in real output prices[34]:

$$\alpha^t = \left(\alpha_L^t \alpha_p^t\right)^{1/2}; \quad t = 1, 2, \ldots. \tag{68}$$

Finally, we look at the problem of defining theoretical indexes for the effects on real income due to changes in real output prices. Define a family of *period t real input quantity growth factors* $\beta\left(x^{t-1}, x^t, p, s\right)$[35]:

$$\beta\left(x^{t-1}, x^t, p, s\right) \equiv g^s\left(p, x^t\right)/g^s\left(p, x^{t-1}\right); \quad s = 1, 2, \ldots. \tag{69}$$

Thus, $\beta\left(x^{t-1}, x^t, p, s\right)$ measures the proportional change in the real income produced by the market sector that is induced by the change in input quantities used by the market sector going from period $t-1$ to t, using the technology that is available during period s and using the reference real output prices p. Thus, each choice of the reference technology s and the reference real output price vector p will generate a possibly different measure of the effect on real income of a change in input quantities going from period t–1 to period t.

Again, it is natural to choose special reference vectors for the measures defined by (69): a *Laspeyres type measure* β_L^t that chooses the period $t-1$ reference technology and reference real output price vector p^{t-1} and a *Paasche type measure* β_P^t that chooses the period t reference technology and reference real output price vector p^t:

[34]The indexes defined by (65)–(67) were defined by Diewert and Morrison (1986, p. 664) in the nominal GDP function context.

[35]This type of index was defined as a true index of value added by Sato (1976, p. 438) and as a real input index by Diewert (1980, p. 456).

$$\begin{aligned}\beta_L^t &\equiv \beta\left(x^{t-1},x^t,p^{t-1},t-1\right)\\ &= g^{t-1}\left(p^{t-1},x^t\right)/g^{t-1}\left(p^{t-1},x^{t-1}\right);\quad t=1,2,\ldots;\end{aligned} \tag{70}$$

$$\beta_P^t \equiv \beta\left(x^{t-1},x^t,p^t,t\right) = g^t\left(p^t,x^t\right)/g^t\left(p^t,x^{t-1}\right);\quad t=1,2,\ldots. \tag{71}$$

Since both measures of real input growth are equally valid, it is natural to average them to obtain an overall measure of the effects of input growth on real income[36]:

$$\beta^t \equiv \left(\beta_L^t\beta_P^t\right)^{1/2};\quad t=1,2,\ldots. \tag{72}$$

Recall that market sector real income for period t was defined by (57) as ρ^t equal to nominal period t factor payments $W^t\cdot x^t$ deflated by the household consumption price deflator P_C^t. It is convenient to define γ^t as the *period t chain rate of growth factor for real income*:

$$\gamma^t \equiv \rho^t/\rho^{t-1};\quad t=1,2,\ldots. \tag{73}$$

It turns out that the definitions for γ^t and the technology, output price and input quantity growth factors $\tau(p,x,t)$, $\alpha\left(p^{t-1},p^t,x,s\right)$, $\beta\left(x^{t-1},x^t,p,s\right)$ defined by (61), (65) and (69), respectively, satisfy some interesting identities, which we will now develop. We have:

$$\begin{aligned}\gamma^t &\equiv \rho^t/\rho^{t-1};\quad t=1,2,\ldots.\\ &= g^t\left(p^t,x^t\right)/g^{t-1}\left(p^{t-1},x^{t-1}\right)\quad \text{using definitions (57)}\\ &= \left[g^t\left(p^t,x^t\right)/g^{t-1}\left(p^t,x^t\right)\right]\left[g^{t-1}\left(p^t,x^t\right)/g^{t-1}\left(p^{t-1},x^t\right)\right]\\ &\quad \left[g^{t-1}\left(p^{t-1},x^t\right)/g^{t-1}\left(p^{t-1},x^{t-1}\right)\right]\\ &= \tau_P^t\alpha\left(p^{t-1},p^t,x^t,t-1\right)\beta_L^t\quad \text{using definitions (63), (65) and (70).}\end{aligned} \tag{74}$$

In a similar fashion, we can establish the following companion identity:

[36]The theoretical indexes defined by (70)–(72) were defined in Diewert and Morrison (1986, p. 665) in the nominal GDP context.

$$\gamma^t \equiv \tau_L^t \alpha\left(p^{t-1}, p^t, x^{t-1}, t\right) \beta_P^t \quad \text{using definitions (62), (65) and (71).} \tag{75}$$

Thus multiplying (74) and (75) together and taking positive square roots of both sides of the resulting identity and using definitions (64) and (72), we obtain the following identity:

$$\gamma^t \equiv \tau^t \left[\alpha\left(p^{t-1}, p^t, x^t, t-1\right) \alpha\left(p^{t-1}, p^t, x^{t-1}, t\right)\right]^{1/2} \beta^t; \quad t = 1, 2, \ldots. \tag{76}$$

In a similar fashion, we can derive the following alternative decomposition for γ^t into growth factors:

$$\gamma^t \equiv t^t \alpha^t \left[\beta\left(x^{t-1}, x^t, p^t, t-1\right) \beta\left(x^{t-1}, x^t, p^{t-1}, t\right)\right]^{1/2}; \quad t = 1, 2, \ldots. \tag{77}$$

It is quite likely that the real output price growth factor $\left[\alpha\left(p^{t-1}, p^t, x^t, t-1\right) \alpha\left(p^{t-1}, p^t, x^{t-1}, t\right)\right]^{1/2}$ is fairly close to α^t defined by (68), and it is quite likely that the input growth factor $\left[\beta\left(x^{t-1}, x^t, p^t, t-1\right) \beta\left(x^{t-1}, x^t, p^{t-1}, t\right)\right]^{1/2}$ is quite close to β^t defined by (72); that is, we have the following approximate equalities:

$$\left[\alpha\left(p^{t-1}, p^t, x^t, t-1\right) \alpha\left(p^{t-1}, p^t, x^{t-1}, t\right)\right]^{1/2} \approx \alpha^t; \quad t = 1, 2, \ldots; \tag{78}$$

$$\left[\beta\left(x^{t-1}, x^t, p^t, t-1\right) \beta\left(x^{t-1}, x^t, p^{t-1}, t\right)\right]^{1/2} \approx \beta^t; \quad t = 1, 2, \ldots. \tag{79}$$

Substituting (78) and (79) into (76) and (77), respectively, leads to the following approximate decompositions for the growth of real income into explanatory factors:

$$\gamma^t \approx \tau^t \alpha^t \beta^t; \quad t = 1, 2, \ldots \tag{80}$$

where τ^t is *a technology growth factor*, α^t is a *growth in real output prices factor* and β^t is a *growth in primary inputs factor*.

Rather than look at explanatory factors for the growth in real market sector income, it is sometimes convenient to express the level of real income in period t in terms of an *index of the technology level* or of Total Factor Productivity in period t, T^t, of the *level of real output prices* in period t, A^t, and of the *level of primary input quantities* in period t, B^t.[37] Thus, we use the growth factors τ^t, α^t and β^t as follows to define the levels T^t, A^t and B^t:

[37] This type of levels presentation of the data is quite instructive when presented in graphical form. It was suggested by Kohli (1990) and used extensively by him (see Kohli 1991, 2003, 2004a, b; Fox and Kohli 1998).

$$T^0 = 1;\ T^t = T^{t-1}\tau^t; \quad t = 1, 2, \ldots; \tag{81}$$

$$A^0 = 1;\ A^t = A^{t-1}\alpha^t; \quad t = 1, 2, \ldots; \tag{82}$$

$$B^0 = 1;\ B^t = B^{t-1}\beta^t; \quad t = 1, 2, \ldots. \tag{83}$$

Using the approximate equalities (80) for the chain links that appear in (81)–(83), we can establish the following approximate relationship for the level of real income in period t, ρ^t, and the period t levels for technology, real output prices and input quantities:

$$\rho^t/\rho^0 \approx T^t A^t B^t; \quad t = 1, 2, \ldots. \tag{84}$$

We now consider a set of assumptions on the technology sets that will ensure that the approximate real income growth decompositions (80) and (84) hold as exact equalities.

Specifically, we follow the example of Diewert and Morrison (1986, p. 663) and assume that the log of the period t (deflated) GDP function, $g^t(p,x)$, has the following *translog functional form*[38]:

$$\begin{aligned} \ln g^t(p,x) \equiv a_0^t + \sum_{m=1}^{M} a_m^t \ln p_m + 1/2 \sum_{m=1}^{M}\sum_{k=1}^{M} a_{mk} \ln p_m \ln p_k \\ + \sum_{n=1}^{M} b_n^t \ln x_n + 1/2 \sum_{n=1}^{M}\sum_{j=1}^{M} b_{nj} \ln x_n \ln x_j \\ + \sum_{m=1}^{M}\sum_{n=1}^{M} c_{mn} \ln p_m \ln x_n; \quad t = 0, 1, 2, \ldots. \end{aligned} \tag{85}$$

Note that the coefficients for the quadratic terms are assumed to be constant over time. The coefficients must satisfy the following restrictions in order for g^t to satisfy the linear homogeneity properties that we have assumed in Sect. 4 above[39]:

[38] This functional form was first suggested by Diewert (1974, p. 139) as a generalization of the translog functional form introduced by Christensen et al. (1971). Diewert (1974, p. 139) indicated that this functional form was flexible.

[39] There are additional restrictions on the parameters which are necessary to ensure that $g^t(p, x)$ is convex in p and concave in x.

$$\sum_{m=1}^{M} a_m^t = 1 \quad \text{for } t = 0, 1, 2, \ldots; \tag{86}$$

$$\sum_{n=1}^{N} b_n^t = 1 \quad \text{for } t = 0, 1, 2, \ldots; \tag{87}$$

$$a_{mk} = a_{km} \quad \text{for all } k, m; \tag{88}$$

$$b_{nj} = b_{jn} \quad \text{for all } n, j; \tag{89}$$

$$\sum_{k=1}^{M} a_{mk} = 0 \quad \text{for } m = 1, \ldots, M; \tag{90}$$

$$\sum_{j=1}^{N} b_{nj} = 0 \quad \text{for } n = 1, \ldots, N; \tag{91}$$

$$\sum_{n=1}^{N} c_{mn} = 0 \quad \text{for } m = 1, \ldots, M; \tag{92}$$

$$\sum_{m=1}^{M} c_{mn} = 0 \quad \text{for } n = 1, \ldots, N. \tag{93}$$

Recall the approximate decomposition of real income growth going from period $t-1$ to t given by (80) above, $\gamma^t \approx \tau^t \alpha^t \beta^t$. Diewert and Morrison (1986, p. 663) showed that if g^{t-1} and g^t are defined by (85)–(93) above, and there is competitive profit-maximizing behaviour on the part of market sector producers for all periods t, then (80) holds as an exact equality[40]; that is, we have

$$\gamma^t = \tau^t \alpha^t \beta^t; \quad t = 1, 2, \ldots. \tag{94}$$

In addition, Diewert and Morrison (1986, pp. 663–665) showed that τ^t, α^t and β^t could be calculated using empirically observable price and quantity data for periods $t-1$ and t as follows:

[40]Diewert and Morrison established their proof using the nominal GDP function $g^t(P, x)$. However, it is easy to rework their proof using the deflated GDP function $g^t(p, x)$ using the fact that $g^t(p, x) = g^t(P/P_C, x) = g^t(P, x)/P_C$ which in turn uses the linear homogeneity property of $g^t(P, x)$ in P.

$$\ln\alpha^t = \sum_{m=1}^{M} 1/2\left(p_m^{t-1}y_m^{t-1}/p^{t-1}y^{t-1} + p_m^t y_m^t/p^t y^t\right)\ln\left(p_m^t/p_m^{t-1}\right) \quad (95)$$
$$= \ln P_T\left(p^{t-1}, p^t, y^{t-1}, y^t\right);$$

$$\ln\beta^t = \sum_{n=1}^{N} 1/2\left(w_n^{t-1}x_n^{t-1}/w^{t-1}x^{t-1} + w_n^t x_n^t/w^t x^t\right)\ln\left(x_n^t/x_n^{t-1}\right) \quad (96)$$
$$= \ln Q_T\left(w^{t-1}, w^t, x^{t-1}, x^t\right);$$

$$\tau^t = \gamma^t/\alpha^t\beta^t \quad (97)$$

where $P_T\left(p^{t-1}, p^t, y^{t-1}, y^t\right)$ is the Törnqvist output price index and $Q_T\left(w^{t-1}, w^t, x^{t-1}, x^t\right)$ is the Törnqvist input quantity index.[41]

Since (80) now hold as exact identities under our present assumptions, Eq. (84), the cumulated counterparts to Eq. (80), will also hold as exact decompositions; that is, under our present assumptions, we have

$$\rho^t/\rho^0 = T^t A^t B^t. \quad t = 1, 2, \ldots. \quad (98)$$

Thus, it is very easy to implement the above decompositions of real income growth into explanatory growth factors, including the *observable measure of technical progress* τ^t defined by the right-hand side of (97), which corresponds to *MFP* growth due to the assumption of constant returns to scale. This result illustrates the exact index number method for estimating productivity growth.[42]

6 Measurement of Output, Labour, Capital and Productivity Indexes in Practice

This section discusses the data needed to implement calculate the index numbers used by NSOs. The key components are an output index and indexes of the primary inputs of labour and capital.

[41]A decomposition of the type in (97) has been used in firm (or more correctly, plant) level analysis to decompose profits; see, e.g., Fox et al. (2003) and Dupont et al. (2005).

[42]For more on this economic approach to index numbers, including dropping the assumptions of perfect competition and constant returns to scale, see Diewert and Fox (2008) and Diewert and Fox (2010). They show that standard index number theory is consistent with quite general cases of imperfect competition. Hence, index number use does not have to be restricted to industries where there is thought to be (close to) perfect competition.

At the aggregate national level, GDP from the national accounts is usually taken as the output measure. In most countries, GDP is calculated using a Laspeyres quantity index. However, due to it being a superlative index, the USA uses a Fisher quantity index. For labour input, hours worked from household labour force surveys are typically used. A simple and popular measure of productivity growth is then GDP growth divided by the growth in aggregate hours worked. This is often used as an indication of the wage growth that can be expected given it represents the growth in real value added produced by workers that exceeds the growth in hours worked. However, some of this growth can be caused by an increase in capital inputs, or capital deepening, hence the interest in *MFP*. With more than one input, an index number formula is used to construct an aggregate input.

Before going into more specifics, it is important to note that many countries, such as the USA, Australia, Canada and the UK, have productivity programs which produce industry-level productivity statistics. The case of Australia is given as an example. Figure 2 plots the cumulated *MFP* indexes for the Australian *Market Sector*,[43] so that the lines represent relative productivity levels compared to the base year, which is the fiscal year 1989–1990.[44] What is immediately striking is the diversity of the productivity growth experience between 1989–1990 and 2017–2018. This illustrates the benefit of industry-level analysis rather than simply a national aggregate approach.[45]

For each industry, value-added growth is taken to be the output series. This is taken from the national accounts and is calculated using a Laspeyres formula. The labour series used can either be a raw hours worked series or a *quality-adjusted labour input* series. This adjusts hours worked to take into account the changing composition of the labour force. The idea is to adjust for improvements in education and for changes in the age and sex distribution of the workforce, reflecting the assumption that differences in wages among types of workers are determined by differences in their productivity. Wage equations are estimated and the predicted wages used in constructing the weights for aggregating over the hours growth of different types of workers. In doing this, the Australian Bureau of Statistics (ABS) follows the approach of the US Bureau of Labor Statistics (see BLS 1993, 2016; Reilly et al. 2005).

[43]There are two versions of the Australian Market Sector; one with twelve industries and one with sixteen industries. Here, the focus is on the twelve original industries. An additional four were added later, and only go back to 1994–1995. Measurement in these additional sectors seems more challenging than the others (such as for the industry "Rental Hiring and Real Estate Services", so are not considered here.

[44]"MFP" is used rather than "TFP" to be consistent with the usual NSO terminology.

[45]For more analysis of the Australian industry level productivity experience, see Fox (2018).

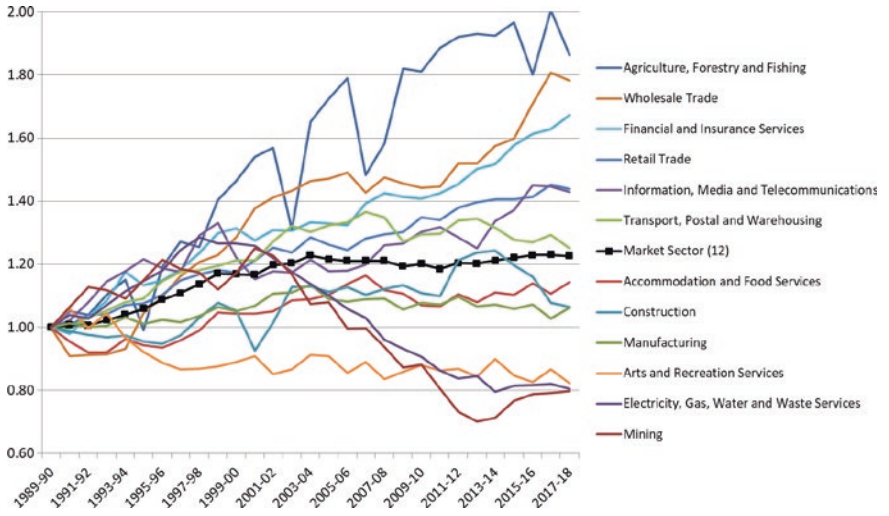

Fig. 2 Multifactor productivity levels, Australian market sector industries (*Source* ABS [2018a]. Note that the indicated years are fiscal years, which run from July 1 to June 30. The plotted series are cumulated indexes, indicating the level of productivity relative to the base year of 1989–1990)

The ABS provides productivity estimates using both quality-adjusted and quality-unadjusted labour inputs. An argument can be made of the use of either. Raw hours worked represent society's time resources dedicated to production. Being able to get more output growth relative to hours worked growth can be interpreted as an enhancement in the use of these resources, which can be interpreted as productivity growth. Alternatively, the fact that society invested in the improvement of labour quality, through, e.g., education, means that more of society's resources are tied up in the hours worked, and hence ignoring that can give a distorted view to the meaning of productivity growth, or improvements in the ability to turn inputs into outputs. Quality change effectively means that the inputs are not the same between periods, and it can be argued that this change needs to be accounted for in productivity measure; this leads to the use of quality-adjusted labour inputs.

For capital, it is too complex to go into detail in this chapter, except to note the following. Estimates of productive capital stocks, based on data on past investment along with estimates of how an asset's services deteriorate over its service life are calculated. Then, it is standard to use a user cost approach for the rental prices of capital (see OECD 2009; Diewert and Fox 2018b). Combined with information from the capital accounts in the national accounts, these rental prices can be used to calculate the cost share of each type of capital considered. The ABS then uses a Törnqvist index to

aggregate over the different types of productive capital to create a *capital services* index. The (quality-adjusted) labour input index and the capital services index are then aggregate, again using a Törnqvist index. Productivity is then defined as the ratio of the value-added index calculated using a Laspeyres index divided by an input index constructed using a Törnqvist index (see ABS 2015a, Chapter 19 for further details).

This mismatch of index number formula between the output and input indexes is not commonly seen in the academic literature, where it is more common to consistently use one index formula for both outputs and inputs, as in the Diewert and Morrison (1986) approach in Sect. 5. The reason why this mismatch occurs in NSO practice is that value added by industry is readily available from the national accounts, where the Laspeyres index tends to be favoured due to being less demanding in terms of data needs (only base period shares are needed for its calculation, whereas the Törnqvist index requires shares from both periods being compared). In addition, the Laspeyres index has an additivity property which is valued in the national accounts community due to it providing a simple additive way of seeing how each component of value added contributes to aggregate growth. From a national accounts point of view, having different outputs indexes for the same industry (depending on the purpose of the index) would be somewhat inconsistent.

For the USA, the numerator of the BLS *MFP* calculation for major sectors is an index of real value added excluding the government, non-profit or household sectors. This is a Fisher index (see BLS 2007, pp. 7–8). The denominator is a Törnqvist quantity index of quality-adjusted labour (BLS 1993) and capital services (BLS 2006).[46] For *MFP* of individual industries, the numerator used by BLS is total gross output for the sector, and the input index is comprised of capital, labour, energy, non-energy materials and purchased business services inputs, where both aggregate output and input indexes have the Törnqvist. That is, intermediate inputs which are subtracted from gross output to produce value added are treated the same as the primary inputs of labour and capital in this case (see Moulton 2018, p. 12). The reason for this approach given by the OECD (2001) is as follows: "At the aggregate level of the economy, gross-output and value-added based measures converge when the gross-output measures are defined as

[46]Thus, there is an inconsistency in the index formula between the numerator and the denominator, due to the numerator being consistent with the use of the Fisher index in calculating GDP in the national accounts. As the Fisher and Törnqvist indexes tend to approximate each other closely in empirical studies, using the Törnqvist index in the numerator is unlikely to make much difference.

sectoral output. Sectoral output is a measure of production corrected for deliveries within a given sector".[47]

As for Australia, the BLS calculates capital services (BLS 2006), and the measure of labour input is adjusted for changes in labour composition in addition to changes in hours worked and uses a Törnqvist index to aggregate over inputs.

Using gross output in the numerator in calculating productivity is often referred to as a *KLEMS* approach, as the input index in the denominator is then comprised of capital (K), labour (L), energy (E), materials (M) and services (S). This approach is synonymous with Dale Jorgenson and his collaborators (see Jorgenson and Timmer 2016).[48]

7 Measurement Challenges

> Why don't we know more after all these years? Our data have always been less than perfect. What is it about the recent situation that has made matters worse? The brief answer is that the economy has changed and that our data-collection efforts have not kept pace with it. "Real" national income accounts were designed in an earlier era, when the economy was simpler…. (Griliches 1994, p. 10)

While productivity slowdowns intensify interest in measurement issues, there remain persistent measurement challenges. New (administrative) data sources and (digital) collection methods can help address these challenges,[49] but the changing nature of the economy presents new challenges or intensifies old ones. Here, we briefly acknowledge some selected challenges and provide references for those interested in potentially contributing to solutions.

---

[47]It can be argued that this explanation is not particularly convincing. Value added-based measures can be considered appropriate for welfare-related issues whereas gross output-based measures are appropriate for issues concerning industrial policy, as gross output is closer to actual enterprise operations than value added. See Balk (2009) and Diewert (2015) for more on the relationship between gross output and value added based measures of productivity growth.

[48]In addition to the headline MFP series published by the ABS (which use value added as the output concept), they also publish experimental gross output-based KLEMS productivity estimates (see ABS 2015b). This requires more effort than calculating the value added-based estimates, as the intermediate inputs have to be "added back" to the value added estimates in the national accounts. Hence, the KLEMS estimates are released with a significant delay relative to the value added-based estimates.

[49]For example, measurement of labour can become more complicated with new occupations emerging and survey respondents being unsure of which industry they are working in. In this case, "administrative" data from employer records can be used to confront the survey results and improve estimates (see ABS 2018b).

Net Output: Depreciation is not a productive asset, yet is part of GDP. It can be argued that depreciation should be subtracted from value added before constructing productivity indexes. This requires removing the depreciation charge in user cost from factor income and subtracting it like an additional intermediate input from the product side (see, e.g., Diewert and Fox 2016). Even if it is agreed that net output is the appropriate output measure, a conceptual issue arises for which there is perhaps no clear resolution yet; Pigou (1941) argued that the key issue was the *maintenance* of physical capital, so only wear-and-tear depreciation should be subtracted from output (or income). This contrasted with the view of Hayek (1941) who argued for the concept of real financial maintenance of capital. This means that asset revaluations represented a decline in income and therefore should also be subtracted along with physical depreciation.

Missing Inputs: Land is often omitted from the list of capital inputs in productivity databases. This is true for the EUKLEMS and World KLEMS databases, which also omit inventories (see Jorgenson and Timmer 2016). When land is included, often extreme assumptions are made (e.g. no change in quantity or quality of the land over time), and the estimates can exhibit concerning patterns (see Alston 2018; Diewert and Fox 2018b).[50]

More broadly, environmental and ecosystem services are typically omitted, even water input for agricultural. This is due to the measurement difficulties of accounting for these inputs. However, some progress continues to be made on improving measurement of these key inputs (see UN 2014a, b). In the productivity context, Brandt et al. (2016) looked at the impact of explicitly accounting for non-renewable resources in productivity measurement, which is a start for a broader economic accounting of natural capital and ecosystem services in productivity measurement. However, several important issues remain unresolved regarding the inclusion of natural

[50]Alston (2018, footnote 9, p. 397): "In evaluating the results from these estimations, I noticed that the USDA-ERS price index for services from land is remarkably volatile, dropping from 1.05 in 1996 to 0.16 in 2000 and 0.12 in 2002 before jumping to 1.35 in 2004. These land rental price gyrations have significant (and seemingly implausible) implications for both the observed and predicted cost share of land (including some negative predicted values from the Translog model) and could well have influenced the cost function estimation results and other analysis using these data. This feature of the land price index appears to be attributable to the practice of treating land as the residual claimant, for the purpose of computing factor payments to land. In their review of the USDA-ERS data, Shumway et al. (2014, 2017) discussed (and largely endorsed) this approach, but they do not appear to have noticed its implications for the measures".

resources and ecosystems in productivity measurement, such as the appropriate method for valuing the services.[51]

There are many more assets that could potentially be considered. For example, Diewert and Fox (2019) advocate having cash balances as part of the asset base.

New Goods and Quality Change: Hulten (2001, p. 29) quoted Adam Smith as follows: "Quality … is so very disputable a matter, that I look upon all information of this kind as somewhat uncertain". However, quality change is an important feature of any modern economy that should not be ignored. NSOs typically try to adjust for quality change using a variety of methods. For example, Landefeld and Grimm (2000) reported that 18% of US final GDP expenditures were deflated using indexes that are calculated with hedonic methods. However, rapid entry of new goods increases the challenge for NSOs. Finding appropriate prices and quantities for goods like cloud computing can be challenging, and appropriately calculating the price declines for the goods early in their lives can be problematic, leading to nominal output deflators being too high and hence downwardly biasing real quantity growth. Much recent measurement effort has gone into dealing with such measurement challenges.[52]

Financial and Insurance Services: This is a very difficult and developing area of measurement. This is problematic for producing aggregate and industry-level productivity growth estimates. Output from Financial and Insurance Services is included in GDP and sectoral productivity estimates tend to be high (see Fig. 2 for Australia). However, debate continues about what exactly this sector does and how its output should be measured (for more on this, see, e.g., Diewert et al. 2016).

Intangibles: A major change in the UN System of National Accounts 2008 was there cognition of expenditure on Research and Development (R&D) as capital formation. Many countries have now implemented this recommendation, along with capitalizing other intangibles such as artistic originals, mineral exploration and computer software. However, there remain other classes of intangibles which could also be capitalized, as investments in these also create assets which lasts more than a period. Due to the pioneering contributions for Corrado et al. (2005, 2006), there are now many studies which calculate estimates of broader classes of intangibles.

[51]See Diewert and Fox (2016) and references therein, Hoang (2018) and Fox et al. (2018).

[52]See, e.g., Bryne et al. (2018), Diewert and Feenstra (2017), Diewert et al. (2018) and references therein.

They classified intangibles into three groups: Computerized Information, Innovative Property and Economic Competencies.[53] See Haskel and Westlake (2017) for an excellent description of the nature, measurement and increasing importance of intangible capital in modern economies.[54]

Digital Economy: As noted in the introduction, the presence of new and free digital goods and services provides significant challenges for measurement. There are broader reasons than the productivity slowdown to suggest that economic statistics are not keeping up with developments in the economy. With the advent of digital cameras, we are taking more photographs than ever; worldwide an estimated 80 billion photos in 2000 and 1.6 trillion in 2015. The price per photo gone from 50 cents to 0 cents, meaning that we are consuming more yet the activity is vanishing from GDP (Varian 2016). Such examples, and the massive increase in consumption of free digital entertainment and communication, raise concerns about the measurement of economic activity and the welfare benefits accruing from the digital economy. Traditional NSO economic statistics are increasingly seen as not reflecting the experiences of businesses and consumers, leading for calls to completely jettison standard frameworks (e.g. Coyle and Mitra-Kahn 2017). Others have advocated extending traditional statistics (Jones and Klenow 2016; Corrado et al. 2017; Heys 2018; Sichel 2019) and improving measurement methodologies to incorporate new data sources (Bean 2016). If measurement is lacking, through methodological challenges, statistical agency budgets or data availability, then we are severely hampered in our ability to understand the impact of new technologies and goods on the economy, and consequently the prospects for future productivity, economic growth and welfare change.

Recent work by Brynjolfsson, Eggers and Gannamaneni (BEG) (2019), Brynjolfsson, Collis, Diewert, Eggers and Fox (BCDEF) (2019) and Diewert et al. (2018, 2019) has opened up new avenues for exploring the

[53]Elnasri and Fox (2017) present results for Australia and examine the implications for productivity from having this broader asset base included in the calculation of the capital services input.

[54]The appropriate way of thinking about asset lives and depreciation of certain intangibles remains an active area of research. For example, Diewert and Huang (2011) proposed an alternative approach to capitalizing R&D.

impact of the digital economy on core economic statistics. BEG (2019) demonstrated that massive online choice experiments can be used to elicit valuations of free digital goods. Specifically, consumers' willingness to accept compensation for losing access to various digital goods can be elicited, providing a valuation. They demonstrated this approach using non-incentive compatible and incentive compatible experiments online, along with laboratory experiments. The incentive compatible choice experiments required participants to potentially give up Facebook for a certain period in exchange for compensation. Their results indicated that digital goods have created large gains in well-being that are missed by conventional measures of GDP and productivity. They concluded that by "periodically querying a large, representative sample of goods and services, including those which are not priced in existing markets, changes in consumer surplus and other new measures of well-being derived from these online choice experiments have the potential for providing cost-effective supplements to existing national income and product".

BCDEF (2019) extended this work to the measurement of welfare change and derived an extended concept of GDP ("GDP-B") which is consistent with this welfare change. This framework provides a means by which to understand the potential mismeasurement that arises from not fully accounting for the digital goods. From running incentive compatible laboratory experiments on the willingness to accept (WTA) to forego consumption for eight popular digital goods, they found that valuations vary dramatically across goods, from a median monthly WTA of over €500 for WhatsApp to €0 for Twitter. Yet the measured prices by NSOs is the same: zero. BCDEF (2019) suggest that a new measure of productivity, Productivity-B, could be calculated, using their extended definition of output, GDP-B.

This literature is still in its infancy. Yet it provides an example of how new data collection approaches, utilizing the reach of the digital economy through online experiments, can be used to enhance our traditional measures of welfare and growth.

8 Conclusion

This chapter has considered productivity theory, measurement and challenges with particular reference to productivity statistics produced by NSOs. As should be clear, the challenges facing NSOs in constructing productivity estimates are not insignificant. This has led to questions about the adequacy

of methods and data for appropriately measuring economic activity and productivity in modern economies.

There are too many complexities (in each of output, capital and labour measurement) to provide a thorough step-by-step guide to NSO practice in a single book chapter, especially given that agencies can follow different approaches. However, the references to the NSO documentation and international manuals, such as those of the OECD (2001, 2009), provide sources for further reading on the range of decisions, methods and data required for producing aggregate and industry productivity statistics.

While the attention here has been on NSO methods, much of material is of course relevant to productivity measurement at the firm level. With the advent of more firm-level databases, constructed using administrative data and using data linking techniques, much more can be learned about macroeconomic performance from examining performance from the firm level upwards (see, e.g., Syverson 2011; Australian Treasury 2018). Combined with new data collection and emerging measurement techniques which take into account the unique features of a digital economy, it can be concluded that this is a very exciting era for productivity measurement.

Acknowledgements The authors thank an anonymous referee for helpful comments, and the editors for their patience. The first author gratefully acknowledges the financial support of the SSHRC of Canada, and both authors gratefully acknowledge the financial support of the Australian Research Council (DP150100830).

References

ABS. 2015a. Australian system of national accounts: Concepts, sources and methods, Cat. No. 5216.0, Australian Bureau of Statistics, Canberra, Australia. https://www.abs.gov.au/ausstats/abs@.nsf/PrimaryMainFeatures/5216.0?OpenDocument.

ABS. 2015b. Experimental estimates of industry level KLEMS Multifactor Productivity, Information Paper, Cat. No. 5260.0.55.003, Australian Bureau of Statistics, Canberra, Australia.

ABS. 2018a. Estimates of industry Multifactor Productivity, 2016–17, Cat. No. 5260.0.55.002, Australian Bureau of Statistics, Canberra, Australia. https://www.abs.gov.au/ausstats/abs@.nsf/mf/5260.0.55.002.

ABS. 2018b. Australian labour account: Concepts, sources and methods, Cat. No. 6150.0, Australian Bureau of Statistics, Canberra, Australia.

Afriat, S.N. 1972. Efficiency estimation of production functions. *International Economic Review* 13: 568–598.

Ahmad, N., and P. Schreyer. 2016. Measuring GDP in a digitalised economy. OECD Statistics Working Papers, 2016/07, OECD Publishing, Paris.

Alston, J.M. 2018. Reflections on agricultural R&D, productivity, and the data constraint: Unfinished business, unsettled issues. *American Journal of Agricultural Economics* 100: 392–413.

Archibald, R.B. 1977. On the theory of industrial price measurement: Output Price indexes. *Annals of Economic and Social Measurement* 6: 57–62.

Australian Treasury. 2018. Firming up productivity in Australia. Presentation to the EMG Workshop 2018, UNSW Sydney, November 29–30. https://treasury.gov.au/productivity/.

Balk, B.M. 1995. Axiomatic price index theory: A survey. *International Statistical Review* 63: 69–93.

Balk, B.M. 1998. *Industrial price, quantity and productivity indices*. Boston: Kluwer Academic.

Balk, B.M. 2003. The residual: On monitoring and benchmarking firms, industries, and economies with respect to productivity. *Journal of Productivity Analysis* 20 (1): 5–47.

Balk, B.M. 2008. *Price and quantity index numbers: Models for measuring aggregate change and difference*. New York: Cambridge University Press.

Balk, B.M. 2009. On the relation between gross output- and value added-based productivity measures: The importance of the Domar factor. *Macroeconomic Dynamics* 13 (Suppl. 2): 241–267.

Basu, S., and J.G. Fernald. 1997. Returns to scale in U.S. production: Estimates and implications. *Journal of Political Economy* 105: 249–283.

Basu, S., and J.G. Fernald. 2002. Aggregate productivity and aggregate technology. *European Economic Review* 46: 963–991.

Bean, C. 2016. Independent review of UK economic statistics. https://www.gov.uk/government/publications/independent-review-of-uk-economic-statistics-final-report.

BLS. 1993. Labor composition and U.S. productivity growth, 1948–90. U.S. Department of Labor, Bureau of Labour Statistics, Bulletin 2426, U.S. Government Printing Office, Washington DC, December.

BLS. 2006. Overview of capital inputs for the BLS Multifactor Productivity measures, July 26. https://www.bls.gov/mfp/mprcaptl.pdf.

BLS. 2007. Technical information about the BLS Multifactor Productivity measures. https://www.bls.gov/mfp/mprtech.pdf.

BLS. 2016. Changes in the composition of labor for BLS Multifactor Productivity measures, 2014. https://www.bls.gov/mfp/mprlabor.pdf.

BLS. n.d. Industry productivity measures. Bureau of Labor Statistics Handbook of Methods. https://www.bls.gov/opub/hom/pdf/homch11.pdf.

Brandt, N., P. Schreyer, and V. Zipperer. 2016. Productivity measurement with natural capital. *Review of Income and Wealth* 63: 7–21.

Bryne, D., C. Corrado, and D. Sichel. 2018. The rise of cloud computing: Minding your P's and Q's. In *Measuring and accounting for innovation in the 21st century*, ed. C. Corrado, J. Miranda, J. Haskel and D. Sichel. NBER Book Series Studies in Income and Wealth, forthcoming.

Brynjolfsson, E., and A. McAfee. 2011. *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Lexington, MA: Digital Frontier Press.

Brynjolfsson, E., and A. McAfee. 2014. *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. New York: W. W. Norton.

Brynjolfsson, E., and J.H. Oh. 2012. The attention economy: Measuring the value of free digital services on the internet. Thirty Third International Conference on Information Systems, Orlando.

Brynjolfsson, E., and A. Saunders. 2009. *Wired for innovation: How information technology is reshaping the economy*. MIT Press.

Brynjolfsson, E., F. Eggers, and A. Gannamaneni. 2019. Using Massive online choice experiments to measure changes in well-being. *PNAS*. www.pnas.org/cgi/doi/10.1073/pnas.1815663116.

Brynjolfsson, E., A. Collis, W.E. Diewert, F. Eggers, and K.J. Fox. 2019. GDP-B: Accounting for the value of new and free goods in the digital economy. NBER Working Paper 25695, Cambridge, MA.

Byrne, D., J. Fernald, and M. Reinsdorf. 2016. Does the United States have a productivity slowdown or a measurement problem? In *Brookings papers on economic activity: Spring 2016*, ed. J. Eberly and J. Stock. Washington, DC: Brookings Institute.

Byushgens (Buscheguennce), S.S. 1925. Obodnomklassegiperpoverkhnostey: popovodu 'idealnovoindeksa' Irving Fischer, a pokupatelnoisilideneg (French translation of Russian title: "Sur uneclasse des hypersurfaces: A propos de 'l'index ideal' de M. Irving Fischer"). *MathematischeskiiSbornik* 32: 625–631.

Caves, D.W., L.R. Christensen, and W.E. Diewert. 1982. The economic theory of index numbers and the measurement of input, output and productivity. *Econometrica* 50: 1393–1414.

Charnes, A., W.W. Cooper, and E. Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2: 429–444.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau. 1971. Conjugate duality and the transcendental logarithmic production function. *Econometrica* 39: 255–256.

Corrado, C., C. Hulten, and D. Sichel. 2005. Measuring capital and technology: An expanded framework. In *Measuring capital in the new economy*, vol. 65, ed. J. Haltiwanger, C. Corrado, and D. Sichel. Studies in Income and Wealth. Chicago: University of Chicago Press.

Corrado, C., C. Hulten, and D. Sichel. 2006. Intangible capital and economic growth. NBER Working Paper no. 11948, National Bureau of Economic Research, Cambridge, MA.

Corrado, C., K.J. Fox, P. Goodridge, J. Haskel, C. Jona-Lasinio, D. Sichel, and S. Westlake. 2017. Improving GDP: Demolishing, repointing or extending? Joint winning entry, Indigo Prize 2017. http://global-perspectives.org.uk/indigo-prize/indigo-prize-winners-2017/.

Cowen, T. 2011. *The great stagnation: How America ate all the low-hanging fruit of modern history, got sick, and will (eventually) feel better*. New York: Dutton.

Coyle, D., and B. Mitra-Kahn. 2017. Making the future count. Joint winning entry, Indigo Prize 2017. http://global-perspectives.org.uk/indigo-prize/indigo-prize-winners-2017/.

Dean, E., and M. Harper. 2001. The BLS productivity measurement program. In *New developments in productivity analysis*, vol. 63, ed. C.R. Hulten, E.R. Dean, and M.J. Harper, 55–84. NBER Studies in Income and Wealth. Chicago: University of Chicago Press.

Diewert, W.E. 1973. Functional forms for profit and transformation functions. *Journal of Economic Theory* 6: 284–316.

Diewert, W.E. 1974. Applications of duality theory. In *Frontiers of quantitative economics*, vol. II, ed. M.D. Intriligator and D.A. Kendrick, 106–171. Amsterdam: North-Holland.

Diewert, W.E. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4: 114–145.

Diewert, W.E. 1978. Superlative index numbers and consistency in aggregation. *Econometrica* 46: 883–900.

Diewert, W.E. 1980. Aggregation problems in the measurement of capital. In *The Measurement of capital*, ed. Dan Usher, 433–528. Chicago: University of Chicago Press.

Diewert, W.E. 1992. Fisher ideal output, input and productivity indexes revisited. *Journal of Productivity Analysis* 3: 211–248.

Diewert, W.E. 1993. The early history of price index research. In *Essays in index number theory*, vol. 1, ed. W.E. Diewert and A.O. Nakamura, 33–65. Amsterdam: North-Holland.

Diewert, W.E. 1997. Commentary on Mathew D. Shapiro and David W. Wilcox: Alternative strategies for aggregating price in the CPI. *The Federal Reserve Bank of St. Louis Review* 79 (3): 127–137.

Diewert, W.E. 2008. Index numbers. In *The new Palgrave dictionary of economics*, 2nd ed, ed. S.N. Durlauf and L.E. Blume. Basingstoke, Hampshire and New York: Palgrave Macmillan.

Diewert, W.E. 2015. Reconciling gross output TFP growth with value added TFP growth. *International Productivity Monitor* 29 (Fall): 60–67.

Diewert, W.E., and R. Feenstra. 2017. Estimating the benefits and costs of new and disappearing products. Discussion Paper 17-10, Vancouver School of Economics, University of British Columbia, Vancouver, BC, Canada.

Diewert, W.E., and K.J. Fox. 1999. Can measurement error explain the productivity paradox? *Canadian Journal of Economics* 32: 251–280.

Diewert, W.E., and K.J. Fox. 2008. On the estimation of returns to scale, technical progress and monopolistic markups. *Journal of Econometrics* 145: 174–193.

Diewert, W.E., and K.J. Fox. 2010. Malmquist and Törnqvist productivity indexes: Returns to scale and technical progress with imperfect competition. *Journal of Economics* 101: 73–95.

Diewert, W.E., and K.J. Fox. 2014. Reference technology sets, free disposal hulls and productivity decompositions. *Economics Letters* 122: 238–242.

Diewert, W.E., and K.J. Fox. 2016. The user cost of non-renewable resources and green accounting. Vancouver School of Economics Discussion Paper 16-01, University of British Columbia.

Diewert, W.E., and K.J. Fox. 2017. Decomposing productivity indexes into explanatory factors. *European Journal of Operational Research* 256: 275–291.

Diewert, W.E., and K.J. Fox. 2018a. Decomposing value added growth into explanatory factors. In *The Oxford handbook of productivity analysis*, ed. E. Grifell-Tatjé, C.A.K. Lovell, and R. Sickles, Chapter 19, 625–662. New York, NY: Oxford University Press.

Diewert, W.E., and K.J. Fox. 2018b. Alternative user costs, productivity and inequality in US business sectors. In *Productivity and inequality*, ed. W.H. Greene, L.A. Khalaf, P. Makdissi, R. Sickles, and M.-C. Voia, Chapter 2, 21–69. Cham: Springer.

Diewert, W.E., and K.J. Fox. 2019. Money and the measurement of total factor productivity. *Journal of Financial Stability* 42: 84–89.

Diewert, W.E., and N. Huang. 2011. Capitalizing R&D expenditures. *Macroeconomic Dynamics* 15: 537–564.

Diewert, W.E., and N.F. Mendoza. 2007. The Le Chatelier principle in data envelopment analysis. In *Aggregation, efficiency, and measurement*, ed. Rolf Färe, Shawna Grosskopf, and Daniel Primont, 63–82. New York: Springer.

Diewert, W.E., and C.J. Morrison. 1986. Adjusting output and productivity indexes for changes in the terms of trade. *The Economic Journal* 96: 659–679.

Diewert, W.E., and A.O. Nakamura. 1999. Benchmarking and the measurement of best practice efficiency: An electricity generation application. *Canadian Journal of Economics* 32: 570–588.

Diewert, W.E., and A.O. Nakamura. 2003. Index number concepts, measures and decompositions of productivity growth. *Journal of Productivity Analysis* 19: 127–159.

Diewert, W.E., and C. Parkan. 1983. Linear programming tests of regularity conditions for production functions. In *Quantitative studies on production and prices*, ed. W. Eichhorn, R. Henn, K. Neumann, and R.W. Shephard, 131–158. Würzburg: Physica-Verlag.

Diewert, W.E., D. Fixler, and K. Zieschang. 2016. Problems with measuring financial services in the national accounts. In *National accounting and economic growth*, ed. J.M. Hartwick. Cheltenham: Edward Elgar.

Diewert, W.E., K.J. Fox, and P. Schreyer. 2018. The digital economy, new products and consumer welfare. ESCoE Discussion Paper 2018-16, Economic Statistics Center of Excellence (ESCoE), London, UK.

Diewert, W.E., K.J. Fox, and P. Schreyer. 2019. Experimental economics and the new goods problem. Vancouver School of Economics Discussion Paper 19-03, University of British Columbia.

Dupont, D.P., K.J. Fox, D.V. Gordon, and R.Q. Grafton. 2005. Profit and price effects of multi-species individual transferable quotas. *Journal of Agricultural Economics* 56: 31–57.

Eichhorn, W. 1978. *Functional equations in economics*. London: Addison-Wesley.

Eichhorn, W., and J. Voeller. 1976. *Theory of the price index*, vol. 140. Lecture Notes in Economics and Mathematical Systems. Berlin: Springer-Verlag.

Elnasri, A., and K.J. Fox. 2017. The contribution of research and innovation to productivity. *Journal of Productivity Analysis* 47: 291–308.

Färe, R. 1988. *Fundamentals of production theory*. Berlin: Springer-Verlag.

Farrell, M.J. 1957. The measurement of production efficiency. *Journal of the Royal Statistical Society, Series A* 120: 253–278.

Feenstra, R. 2004. *Advanced international trade: Theory and evidence*. Princeton, NJ: Princeton University Press.

Feldstein, M. 2017. Understanding the real growth of GDP, personal income, and productivity. *Journal of Economic Perspectives* 31: 145–164.

Fisher, I. 1911. *The purchasing power of money*. London: Macmillan.

Fisher, I. 1922. *The making of index numbers*. Boston: Houghton-Mifflin.

Fisher, F.M., and K. Shell. 1972. The pure theory of the national output deflator. In *The economic theory of price indexes*, 49–113. New York: Academic Press.

Fox, K.J., and U. Kohli. 1998. GDP Growth, terms of trade effects and total factor productivity. *Journal of International Trade and Economic Development* 7: 87–110.

Fox, K.J., R.Q. Grafton, J. Kirkley, and D. Squires. 2003. Property rights in a fishery: Regulatory change and firm performance. *Journal of Environmental Economics and Management* 46: 156–177.

Fox, K.J., K.V. Hoang, and S. Zeng. 2018. Value added and productivity decompositions with natural capital. Presented at the ESCoE Conference on Economic Measurement, Bank of England, May 16–17.

Frisch, Ragnar. 1930. Necessary and sufficient conditions regarding the form of an index number which shall meet certain of fisher's tests. *Journal of the American Statistical Association* 25 (172): 397–406.

Frisch, R. 1936. Annual survey of general economic theory: The problem of index numbers. *Econometrica* 4: 1–39.

Funke, H., and J. Voeller. 1978. A note on the characterisation of fisher's ideal index. In *Theory and applications of economic indices*, ed. W. Eichhorn, R. Henn, O. Opitz, and R.W. Shephard, 177–181. Würzburg: Physica-Verlag.

Funke, H., and J. Voeller. 1979. Characterization of fisher's ideal index by three reversal tests. *StatistischeHefte* 20: 54–60.

Gordon, R. 2016. *The rise and fall of American growth: The U.S. standard of living since the Civil War*. Princeton, NJ: Princeton University Press.

Gorman, W.M. 1968. Measuring the quantities of fixed factors. In *Value, capital and growth: papers in honour of Sir John Hicks*, ed. J.N Wolfe, 141–172. Chicago: Aldine Press.

Greenstein, Shane, and Ryan C. McDevitt. 2011. The broadband bonus: Estimating broadband internet's economic value. *Telecommunications Policy* 35 (7): 617–632.

Griliches, Z. 1994. Productivity, R&D, and the data constraint. *American Economic Review* 84: 1–23.

Groshen, E.L., B.C. Moyer, A.M. Aizcorbe, R. Bradley, and D.M. Friedman. 2017. How government statistics adjust for potential biases from quality change and new goods in an age of digital technologies; A view from the trenches. *Journal of Economic Perspectives* 31 (2): 187–210.

Haskel, J., and S. Westlake. 2017. *Capitalism without capital: The rise of the intangible economy*. Princeton and Oxford: Princeton University Press.

Heys, R. 2018. Challenges in measuring the modern economy. Presentation at the ESCoE Conference, Bank of England, London, 16–17 May.

Hoang, K.V. 2018. Accounting for natural capital in mining MFP: Comparing user costs for non-renewable resources. Paper presented at the IARIW General Conference, Copenhagen, 25–30 August.

Hotelling, H. 1932. Edgeworth's taxation paradox and the nature of demand and supply functions. *Journal of Political Economy* 40: 577–616.

Hulten, C.R. 2001. Total factor productivity: A short biography. In *New developments in productivity analysis*, ed. C.R. Hulten, E.R. Dean, and M.J. Harper, 1–54., NBER Chicago: University of Chicago Press.

Hulten, C.R., and L. Nakamura. 2017. We see the digital revolution everywhere but in GDP. Presentation to the NBER/CRIW conference on Measuring and Accounting for Innovation in the 21st Century, Washington, DC, March 10. http://conference.nber.org/confer/2017/CRIWs17/program.html. Accessed March 10, 2017.

IMF. 2018. Measuring the digital economy. Staff Report, February. http://www.imf.org/external/pp/ppindex.aspx.

Jones, C.I., and P.J. Klenow. 2016. Beyond GDP? Welfare across countries and time. *American Economic Review* 106: 2426–2457.

Jorgenson, D.W., and Z. Griliches. 1967. The explanation of productivity change. *Review of Economic Studies* 34: 249–283.

Jorgenson, D.W., and Z. Griliches. 1972. Issues of growth accounting: A reply to Edward F. Denison. *Survey of Current Business* 55 (5): part II, 65–94.

Jorgenson, D.W., and M. Nishimizu. 1978. U.S. and Japanese economic growth, 1952–1974. *Economic Journal* 88: 707–726.

Jorgenson, D.W., and M. Timmer. 2016. *World KLEMS*. Harvard University and the University of Groningen. http://www.worldklems.net/index.htm.

Kohli, U. 1978. A gross national product function and the derived demand for imports and supply of exports. *Canadian Journal of Economics* 11: 167–182.

Kohli, U. 1990. Growth accounting in the open economy: Parametric and nonparametric estimates. *Journal of Economic and Social Measurement* 16: 125–136.

Kohli, U. 1991. *Technology, duality, and foreign trade: The GNP function approach to modeling imports and exports*. New York: Harvester Wheatsheaf.

Kohli, Ulrich. 2003. GDP growth accounting: A national income function approach. *Review of Income and Wealth* 49 (1): 23–34.

Kohli, U. 2004a. An implicit Törnqvist index of real GDP. *Journal of Productivity Analysis* 21: 337–353.

Kohli, U. 2004b. Real GDP, real domestic income, and terms-of-trade changes. *Journal of International Economics* 62: 83–106.

Konüs, A.A. 1939. The problem of the true index of the cost of living. *Econometrica* 7: 10–29.

Konüs, A.A., and S.S. Byushgens. 1926. K problemepokupatelnoicilideneg (English translation of Russian title: On the problem of the purchasing power of money). *VoprosiKonyunkturi* II (1) (supplement to the Economic Bulletin of the Conjuncture Institute): 151–171.

Landefeld, J.S., and B. Grimm. 2000. A note on the impact of hedonics and computers on real GDP. *Survey of Current Business*, December 17–22.

Laspeyres, E. 1871. Die BerechnungeinermittlerenWaarenpreissteigerung. *Jahrbücher fur Nationalökonomic und Statistik* 16: 296–314.

Lau, L.J. 1976. A characterization of the normalized restricted profit function. *Journal of Economic Theory* 12: 131–163.

Lawrence, D., and W.E. Diewert. 2006. Regulating electricity networks: The ABC of setting X in New Zealand. In *Performance measurement and regulation of network utilities*, ed. T. Coelli and D. Lawrence, 207–241. Cheltenham: Edward Elgar.

McFadden, D. 1978. Cost, revenue and profit functions. In *Production economics: A dual approach to theory and applications*, vol. 1, ed. M. Fuss and D. McFadden, 3–109. Amsterdam: North-Holland.

Mokyr, J., C. Vickers, and N.L. Ziebarth. 2015. The history of technological anxiety and the future of economic growth: Is this time different? *Journal of Economic Perspectives* 29 (3): 31–50.

Moulton, B. 2018. The measurement of output, prices, and productivity: What's changed since the Boskin Commission? The Brookings Institution, July 25. https://www.brookings.edu/research/the-measurement-of-output-prices-and-productivity/.

Nakajima, T., M. Nakamura, and K. Yoshioka. 1998. An index number method for estimating scale economies and technical progress using time-series of cross-section data: Sources of total factor productivity growth for Japanese manufacturing, 1964–1988. *Japanese Economic Review* 49: 310–334.

Nakajima, T., A. Nakamura, and M. Nakamura. 2002. Japanese TFP growth before and after the financial bubble: Japanese manufacturing industries. Paper presented at the NBER, Cambridge, MA, July 26.

Norman, R.G., and S. Bahiri. 1972. *Productivity measurement and incentives*. Oxford: Butterworth-Heinemann.

OECD. 2001. *Measuring productivity—OECD manual*: *Measurement of aggregate and industry-level productivity growth*. Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264194519-en.

OECD. 2009. *Measuring capital—OECD manual*, 2nd ed. Paris: OECD Publishing.

OECD. 2018. OECD compendium of productivity indicators 2018. Paris: OECD Publishing. http://dx.doi.org/10.1787/pdtvy-2018-en.

Paasche, H. 1874. Über die Preisentwicklung der letztenJahrenach den Hamburger Borsennotirungen. *Jahrbücher fur Nationalökonomie und Statistik* 12: 168–178.

Pigou, A.C. 1941. Maintaining capital intact. *Economica* 8: 271–275.

Reilly, R., W. Milne, and S. Zhao. 2005. Quality-adjusted labour inputs. Research Paper (1351.0.55.010), Australian Bureau of Statistics, Canberra, Australia.

Salter, W. E. G. 1960. *Productivity and technical change*. Cambridge: Cambridge University Press.

Samuelson, P.A. 1953. Prices of factors and goods in general equilibrium. *Review of Economic Studies* 21: 1–20.

Samuelson, P.A., and S. Swamy. 1974. Invariant economic index numbers and canonical duality: Survey and synthesis. *American Economic Review* 64: 566–593.

Sato, K. 1976. The meaning and measurement of the real value added index. *Review of Economics and Statistics* 58: 434–442.

Shumway, C.R., B.M. Fraumeni, L.E. Fulginiti, J.D. Samuels and S.E. Stefanou. 2014. *Measurement of U.S. agricultural productivity: A 2014 review of current statistics and proposals for change*. School of Economic Science, Washington State University (Reproduced from the Report of the Agricultural Productivity Committee to the USDA Economic Research Service).

Shumway, C. Richard, Barbara M. Fraumeni, Lilyan E. Fulginiti, Jon D. Samuels, Spiro E. Stefanou. 2017. U.S. Agricultural Productivity: A Review of USDA Economic Research Service Methods. *Applied Economic Perspectives and Policy*: ppv032.

Sichel, D. 2016. Two books for the price of one: Review article of *The rise and fall of American growth* by Robert J. Gordon. *International Productivity Monitor* 31 (Fall): 57–62.

Sichel, D. 2019. Productivity measurement: Racing to keep up. NBER Working Paper 25558, Cambridge, MA.

Syverson, C. 2011. What determines productivity? *Journal of Economic Literature* 49: 326–365.

Syverson, C. 2017. Challenges to mismeasurement explanations for the U.S. productivity slowdown. *Journal of Economic Perspectives* 31: 165–186.

Törnqvist, L. 1936. The Bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin* 10: 1–8.

UN/EU/FAO/OECD/World Bank. 2014a. *System of environmental-economic accounting: Central framework*. New York: United Nations. http://unstats.un.org/unsd/envaccounting/seeaRev/SEEA_CF_Final_en.pdf.

UN/EU/FAO/OECD/World Bank. 2014b. *System of environmental-economic accounting: Experimental ecosystem accounting*. New York: United Nations. http://unstats.un.org/unsd/envaccounting/seeaRev/eea_final_en.pdf.

Varian, H.R. 1984. The nonparametric approach to production analysis. *Econometrica* 52: 579–597.

Varian, H. 2016. A microeconomist looks at productivity: A view from the valley. Presentation, Brookings. https://www.brookings.edu/wpcontent/uploads/2016/08/varian.pdf.

von Hayek, F.A. 1941. Maintaining capital intact: A reply. *Economica* 8: 276–280.

Walsh, C.M. 1901. *The measurement of general exchange value*. New York: Macmillan.

Walsh, C.M. 1921a. *The problem of estimation*. London: P.S. King & Son.

Walsh, C.M. 1921b. Discussion. *Journal of the American Statistical Association* 17: 537–544.

Woodland, A.D. 1982. *International trade and resource allocation*. Amsterdam: North-Holland.